DOCUMENT RESUME

ED 157 918 95 TM 007 255

AUTHOR Weisberg, Herbert I; Haney, Walt
TITLE Longitudinal Evaluation of Head Start Planned Variation and Follow Through.
INSTITUTION Huron Inst., Cambridge, Mass.
SPONS AGENCY Bureau of Elementary and Secondary Education (DHEW/OE), Washington, D.C. Div. of Follow Through.
PUB DATE Jun 77
CONTRACT OEC-0-74-0394
NOTE 108p.

EDRS PRICE MF-$0.83 HC-$6.01 Plus Postage.
DESCRIPTORS Academic Achievement; Achievement Gains; Compensatory Education Programs; *Data Analysis; Disadvantaged Youth; Elementary Education; Evaluation Methods; Followup Studies; *Longitudinal Studies; *Program Evaluation; Research Design; *Research Problems; *Statistical Bias; Statistical Data
IDENTIFIERS *Project Follow Through; *Project Head Start

ABSTRACT

Both administratively and in terms of their evaluations, Head Start and Follow Through have been treated as separate programs. Follow Through was initially conceived, however, as an effort to preserve and augment Head Start gains. In this report, achievement test data on a set of children who participated in both Head Start and Follow Through are used to determine if Follow Through helps maintain the benefits of Head Start in the early elementary grades, and how the Head Start experience of children may have confounded efforts in the national evaluation of Follow Through to calculate program effects. Regarding the maintenance of benefits, Follow Through does maintain increased test scores on skills measured through the end of kindergarten. There are, however, no overall effects on skills measured in later grades. As for the question of measurement problems, there is probably a bias operating against Follow Through in general, and particularly in sites where the Follow Through group has a much higher proportion of Head Start children than the Non-Follow Through group. This bias is not overwhelming, but may be large enough to change the outcomes significantly for many sites. In addition, the design for data collection in Follow Through severely limits the ability to investigate how Follow Through has affected children's growth. Ignoring Head Start, and using one pre-test and one post-test forces a great burden on statistical adjustments. In retrospect, it would have been preferable to rely less on statistical manipulations of the data and more on tracing data for individual children across several time points.
(Author/CP)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED157918

SCOPE OF INTEREST NOTICE
The ERIC Facility has assigned this document for processing to: TM PS
In our judgement, this document is also of interest to the clearinghouses noted to the right. Indexing should reflect their special points of view.

Longitudinal Evaluation of Head Start Planned Variation and Follow Through

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Herbert I. Weisberg

Walt Haney

June 1977

Huron Institute
123 Mt. Auburn Street
Cambridge, Massachusetts 02138

Phone: 617/661-9285

Prepared for:
The Office of Education
Follow Through Evaluation Branch
Contract No. OEC-0-74-0394

TM007 255

Table of Contents

		Page
I.	Introduction and Background	1
II.	Head Start-Follow Through Data and the Effects of Follow Through in Maintaining the Benefits of Head Start	4
	- Acquisition of the Data	7
	- General Overview of the Data	11
	- Follow Through Effects on Head Start	17
III.	Confounding Problems in Estimating Average Follow Through Effects	24
	- Confounding in the National Evaluation	25
	- Conceptual Framework	28
IV.	Analysis of the Data	35
	- Discussion of Tables	39
V.	Conclusion	55
	- Limitations of the Data	55
	- Implications for Confounding in the National Evaluation	57
	- A Rough Quantitative Estimate of the Bias in the National Evaluation	65
	- Implications for Future Studies	70
	Bibliography	
	Appendix A: Head Start Planned Variation-Follow Through Data Summaries (by cohort, site and treatment group)	
	Appendix B: Merger of HSPV and FT Test Data	

Tables and Figures

		Page
Table 1:	Maximum Sample Sizes for Each of Four Treatment Groups by Site and Cohort	14
Table 2:	Head Start Data for Original HSPV Sample, HS-FT and HS-NFT Groups, Cohort III	18
Table 3:	Head Start Data for Original HSPV Sample, HS-FT and HS-NFT Groups, Cohort IV	19
Table 4:	Cohort IV Traced HS Children: HS-FT versus HS-NFT. READING	22
Table 5:	Cohort III. Blocked NPS-NFT Group, Traced Reading Scores	41
Table 6:	Cohort III. Blocked NPS-FT Group, Traced Reading Scores	42
Table 7:	Cohort III. Blocked HS-FT Group, Traced Reading Scores	44
Table 8:	Cohort III. Blocked NPS-NFT Group, Traced Math Scores	48
Table 9:	Cohort III. Blocked NPS-FT Group, Traced Math Scores	49
Table 10:	Cohort III. Blocked HS-FT Group, Traced Math Scores	50
Table 11:	Cohort IV. Blocked HS-NFT Group, Traced Reading Scores	51
Table 12:	Cohort IV. Blocked HS-FT Group, Traced Reading Scores	52
Table 13:	Cohort IV. Blocked HS-NFT Group, Traced Math Scores	53
Table 14:	Cohort IV. Blocked HS-FT Group, Traced Math Scores	54
Table 15:	Estimated Biases Associated with Preschool Confounding	67
Table 16:	Patterns of Site Effects by Sponsor with and without Preschool Correction	69
Figure 1:	Relationships between HSPV and FT Projects	6
Figure 2:	Testing Patterns for HSPV Longitudinal Evaluation	8
Figure 3:	Tests Administered to HSPV-FT Cohorts III and IV	9

Abstract

Both administratively and in terms of their evaluations, Head Start and Follow Through have been treated as separate programs. Follow Through was initially conceived, however, as an effort to preserve and augment Head Start gains. In this report we analyze achievement test data on a set of children who participated in both Head Start and Follow Through. We focus on two main issues:

Whether Follow Through helps maintain the benefits of Head Start in the early elementary grades;

The way in which the Head Start experience of children may have confounded efforts in the national evaluation of Follow Through to calculate program effects.

Regarding the first, we conclude that Follow Through does appear to maintain increased test scores on skills measured through the end of kindergarten. There are, however, no overall effects on skills measured in later grades.

Regarding the second, we conclude that there is probably a bias operating against Follow Through in general, and particularly in sites where the Follow Through group has a much higher proportion of Head Start children than the Non-Follow Through group. This bias is not overwhelming, but may be large enough to change the outcomes significantly for many sites.

In addition, we note that the design for data collection in Follow Through severely limits our ability to investigate how Follow Through has affected children's growth. Ignoring Head Start, and using one pre-test and one post-test forces a great burden on statistical adjustments. In retrospect, we think that it would have been preferable to rely less on statistical manipulations of the data and more on tracing data for individual children across several time points.

I. Introduction and Background

Head Start was begun during a flush of enthusiasm in the mid-1960's for massive programs to ameliorate social problems. It was well accepted by then that poor children started school comparatively disadvantaged in terms of their ability to progress rapidly in the attainment of cognitive skills. Moreover, faith in planned intervention to promote intellectual development was at its height. There was considerable optimism that children's potential for cognitive growth could be substantially influenced by providing an enriched environment.

In part Head Start was predicated on the hope that a limited intervention over a period of less than a year prior to school entry would compensate for lack of opportunities up to that time. It was hoped that a summer-long or year-long compensatory program would give children the "head start" they needed to start off school on an equal footing with middle class children and progress from there on at a comparable rate. This optimistic "take-off" hypothesis posited that a child's school entry level could be boosted by a special preschool program and he or she would take off from this new level upon entering school.

Several early studies of the effects of Head Start on achievement test and I.Q. scores quickly deflated these expectations. A well-publicized small study by Wolff and Stein (1966) suggested that gains in academic achievement resulting from Head Start faded out even before children finished kindergarten. Also, the Westinghouse-Ohio national evaluation of Head Start (Cicirelli, 1969) indicated that after the first year or two of school, children with Head Start experience displayed growth trajectories quite similar to other disadvantaged children without Head

Start. Let us refer to this description of the effect of Head Start as "fade-out".

The evaluations which suggested fade-out were subject to serious methodological flaws (c.f. Campbell-Ehrlebacher, 1970; Smith and Bissell, 1970). So it was not clear whether a complete fade-out is an accurate description of what actually happened to the effects of Head Start. Nevertheless, the early evaluations of Head Start seemed to make one thing clear: the hoped-for take-off was not occurring.

This sort of conclusion was interpreted by many as a sign that compensatory education was a failure in terms of its ability to promote cognitive development (Mendelsohn, 1970, p. 448). Others felt that the logic of compensatory education was sound, but that Head Start was not an extensive enough intervention. There was a need to "follow through" on the preschool preparation in order to achieve a lasting effect. The expected take-off would occur later, after several years of sustained intervention.

Project Follow Through was motivated in large part by this rationale. In fact, when President Johnson first proposed Follow Through he reasoned that "we should strengthen the Head Start program, begin it for children 3 years old, and maintain its educational momentum by following through in the early years of school." One month later, he asked Congress to "preserve the hope and opportunity of Head Start by a 'followthrough' program," adding that "to fulfill the rights of America's children to equality of educational opportunity, the benefits of Head Start must be carried to the early grades" (Elmore, 1976, p. 112).

Despite the notion that Project Follow Through would carry the benefits of Head Start into the early grades of elementary school, the relationship between the two programs has received little attention in the massive evaluation literature concerning the Head Start and Follow Through programs. In fact, there exists very little data of any kind linking the performance of children in Head Start and in Follow Through. One part of the Follow Through national evaluation data does, however, provide the opportunity to examine the Follow Through performance of a small set of children who were previously tested in a special Head Start program. The purpose of this paper is to illuminate the Head Start-Follow Through connection, by analyzing this unique data set.

More specifically, our purposes are twofold:

--To attempt an evaluation of whether Follow Through helps maintain the benefits of Head Start into the early grades of elementary school.

--To estimate the degree to which the Head Start experience of Follow Through students may confound efforts in the national evaluation of Follow Through to calculate program effects.

The report has five chapters. In this first chapter we have briefly recounted some of the background to the Head Start and Follow Through programs and some of the assumptions about children's growth implicit in the ideas of "head start" and "follow through." The second chapter then describes the Head Start-Follow Through data set employed in this study and our attempts to compare the performance of Head Start graduates who attended Follow Through with others who did not attend Follow Through. The third chapter describes

the way in which the Head Start experience of Follow Through children may confound the estimation of Follow Through program effects and presents a conceptual model which helps to clarify the confounding problem. Chapter IV presents the analyses we conducted in an empirical examination of the Head Start confounding problem and chapter V the conclusions drawn from our analyses.

II. The Head Start-Follow Through Data and the Effects of Follow Through in Maintaining the Benefits of Head Start

The data we have analyzed come from a special sort of Head Start program, namely the Head Start Planned Variation program. From 1969 to 1972 three waves or cohorts of children attended Head Start programs which employed curriculum models developed and employed in the Follow Through program. These programs were all located in sites which also had Follow Through programs using the same curriculum models. In 1969-70 the HSPV program employed eight curriculum models in sixteen sites, in 1970-71 twelve models in 37 sites, and in 1971-72 twelve models in 36 sites. Separate evaluations have previously been conducted of each of the three HSPV waves or cohorts of children (Dittman, et al., 1971, Smith, 1973; Weisberg, 1974).

After graduating from their respective preschool programs, some of the HSPV children entered the Follow Through program. The relationships between the three HSPV cohorts and the Follow Through evaluation cohorts are illustrated in Figure 1.[1] For a variety of reasons the continuity of testing of children in the HSPV-FT sites was less than ideal in the early years, and consequently a large number of children in the HSPV sites were not included in the tested sample. This led to a joint decision between the Office of Child Development and the Office of Education to restrict the HSPV-FT test sample to seven projects in the third wave of Follow Through and to thirteen projects in the fourth FT wave, and to eliminate Follow Through entry-level testing as a criterion for inclusion in the testing of HSPV-FT children (Smith, 1972; David, 1973). Also, this decision restricted the HSPV-FT testing solely to sites in which children entered kindergarten instead of entering directly into first grade. The HSPV-FT sites in FT CIII have been tested at seven points. The twelve HSPV-FT sites in FT CIV have, however, been tested at only five points. They were not tested upon entering kindergarten because they were not then included in the FT test sample and were not to be tested upon leaving

---

[1] Note that the second HSPV cohort went into the third FT cohort and the third HSPV cohort into FT cohort IV. To avoid confusion we hereafter refer to the groups analyzed in this report by the FT cohort labels, CIII and CIV.

Figure 1

Relationships Between Head Start Planned Variation and Follow Through Projects

FT Cohorts*	1969-70	1970-71	1971-72	1972-73	1973-74	1974-75	1975-76
I	K 1	1 2	2 3	3			
II	HSPVCIV	K 1	1 2	2 3	3		
III		HSPVCII°	K° 1	1° 2	2° 3	3°	
IV			HSPVCIII°	K° 1	1° 2	2° 3	3

*Each FT cohort contains two streams, an entering kindergarten (EK) stream and an entering first grade (E1) stream. The HSPV sites have been coordinated only with the EK streams of FT. Note also that FT cohort II corresponds to HSPV cohort I (HSPVCI) and FT cohort III to HSPVCII, etc.

°Cohorts from which data are analyzed in this report.

third grade because FT testing was curtailed for economy's sake. The tests administered to each cohort at each of these test points are summarized in Figure 3.

Acquisition of Data

After determining potential patterns of test availability as outlined above, we began negotiations in the summer of 1975 to obtain Follow Through test data for children in each of the HSPV-FT sites. Since the Huron Institute conducted the evalautions of HSPV, the preschool data for the relevant sites were already available to us.

Initially, it seemed as if Abt Associates (the organization conducting analyses of the National Follow Through evaluation data) would be the most convenient source of the needed FT test data. However, by mid-September 1975, it had become apparent that there would be a number of difficulties in our obtaining the necessary Follow Through data from Abt--even though Abt personnel were extending to us all assistance possible. The problems were:

1) Abt did not at that time possess all of the data we needed.
2) The Abt CDC computing system was not easily compatible with the IBM computers at Harvard University which Huron normally uses.
3) Since Abt had no mandate from OE to work with CIV data, these data had not been "cleaned" after their arrival from SRI. Hence, the cost of working up these data to fulfill our requests would be very high.

Figure 2: Testing Patterns for HSPV Longitudinal Evaluation

			Cohort, School Year, Grade Level, Fall or Spring													
			FT Cohort III							FT Cohort IV						
			1970-71		1971-72		1972-73		1973-74	1974-75	1971-72		1972-73		1973-74	1974-7
			HeadStart		K		1		2	3	HS		K		1	2
FT Sponsor	Sites		F	S	F	S	S		S	S	F	S	F	S	S	S
Far West	Duluth	0204	X	X	X	X	X		X	X	X	X		X	X	X
	Salt Lake City	0209									X	X		X	X	X
	Tacoma, WA	0213	X	X	X	X	X		X	X	X	X		X	X	X
Arizona	Lakewood, NJ	0309									X	X		X	X	X
	Lincoln, NE	0316	X	X	X	X	X		X	X	X	X		X	X	X
Bank Street	Wilmington, DE	0511									X	X		X	X	X
Kansas	Portageville, MO	0804.	X	X	X	X	X		X	X	X	X		X	X	X
	Mounds, IL	0808									X	X		X	X	X
High/Scope	Oskaloosa Co, FL	0902	X	X	X	X	X		X	X	X	X		X	X	X
	Greeley, CO	0906	X	X	X	X	X		X	X	X	X		X	X	X
	Seattle, WZ	0910	X	X	X	X	X		X	X						
Pittsburgh	Lock Haven, PA	1203									X	X		X	X	X
	Montevideo, MN										X	X		X	X	X

Figure 3: Tests Administered to HSPV-FT Cohorts III and IV

	Head Start		Kindergarten		First Grade	Second Grade	Third Grade
	Fall	Spring	Fall	Spring	Spring	Spring	Spring
HSPV-FT	PSI (64-item)	PSI (64-item)	PSI (29-item)		MAT I	MAT II	MAT Elementary
Cohort III	NYU Booklets 3D & 4A	NYU Booklets 3D & 4A	WRAT (SRI form)	WRAT (SRI form)			Ravens
(HSPV CII)	Stanford-Binet (admin. to random half)	Stanford-Binet	PPVT (SRI modified Form A)	PPVT (SRI modified Form A)			Coopersmith
	Motor-Inhibition	Motor-Inhibition	Lee-Clark	Locus-of-Control			IAR Scale
	CPSCS		MAT Primer (Form F)				
HSPV-FT	PPVT	PPVT	None	MAT Primer (Form F)	MAT I	MAT II	None
Cohort IV	PSI (32-item) WRAT	PSI (32-item) WRAT					
(HSPV CIII)	Classroom Behavior Inventory	Classroom Behavior Inventory					
	Brown-Self-Concept	Gumpgookies					
	ETS Enumeration	ETS Enumeration					
	ITPA Motor-Inhibition	ITPA Motor-Inhibition					
	8-Block Sort Test	8-Block Sort Test					
		Rel. Redun. Cue					

In light of these problems, we decided that it would be far more efficient for us to obtain data directly from the Stanford Research Institute. In early November, with the assistance of Richard Marciano of SRI, we submitted a detailed specification of the data needed from SRI.

A five-month delay resulted from uncertainty over the conditions under which SRI could legally provide Huron with the necessary data without violating Federal regulations concerning confidentiality of such data. This problem was not resolved until February 1976, when the general counsel of the Department of Health, Education and Welfare rendered an opinion that SRI could legally transmit data to "organizations conducting studies for, or on behalf of, educational agencies or institutions for the purpose of... improving instruction" (Education Division, 1976).

After this issue was resolved, SRI was able to meet our requests and on March 22 we obtained three data tapes containing the requested information. Descriptions of these data tapes can be found in the data file description entitled "Data-Sharing, Huron Institute, Head Start Planned Variation Projects Within Follow Through Evaluation, School Years 1971/72 Through 1974/75", dated March 19, 1976.

From these data tapes, and the tapes containing the HSPV data already in Huron's possession, we created a merged file for each of the two cohorts. These files included all children whose FT and HSPV records could be reliably matched. In the merged files, each child's record contains

information on Follow Through experience, background information collected during Follow Through, test scores during Follow Through, Head Start experience, background information collected at Head Start entry, and test scores during Head Start. In addition, we thought the standard score transformation provided by Harcourt Brace for the Metropolitan Achievement Test (MAT) might provide a more meaningful index for describing growth. Consequently, we have constructed a record containing the standard scores for the MAT reading and mathematics subtests and incorporated this into the file along with raw scores.[2]

To our knowledge, these files represent the only data set with information on the experience and development of children from HS entry through the end of third grade. While it is in many respects painfully limited, it represents a unique resource which required a considerable effort to create and may be of interest for purposes of secondary analysis.

General Overview of the Data

While many aspects of our data set are of potential interest, we are primarily concerned with the way Head Start and Follow Through interactively affect growth over time. In focusing on this longitudinal aspect of the data, we decided to limit our attention to the principal academic achievement measures used since neither the non-cognitive measures nor general cognitive problem-solving measures were administered at interim points mid-way through the Follow Through program (i.e., end of first or second grade).

---

[2] The process of merging the HSPV and FT data sets is described in Appendix B.

From our previous experience in analyzing HSPV data, we concluded that the Pre-School Inventory (PSI) was the most meaningful and reliable index of "academic" readiness (Weisberg, 1974). Unfortunately, three different versions of the PSI appear in our data. A 64-item version was used in Cohort III HS, a 29-item version at fall kindergarten (FK) for Cohort III, and a 32-item in Cohort IV HS (i.e., HSPV CIII). While this inconsistency creates a problem, we have no viable alternative but to use these different forms. Note from Figure 1 that none of the tests administered was given in both HS and FT for a given cohort, and none was given in HS to both cohorts. Thus there is no consistency of metric between HSPV and FT, and the only consistency across cohorts is the use of the MAT. The MAT was administered in Cohort III at fall kindergarten (FK), spring kindergarten (SK), spring of first grade (S1), spring of second grade (S2), and spring third (S3). In cohort IV, there were no entry (FK) or exit (S3) data, but intermediate testing (SK, S1, and S2) did take place.

The Head Start-Follow Through data are presented in summary form in Appendix A. These data are divided into the four relevant treatment groups for each cohort; the Non-Preschool-Non-Follow Through (NPS-NFT), the Non-Preschool-Follow Through children (NPS-FT), the Head Start-Non-Follow Through children (HS-NFT) and the Head Start-Follow Through children (HS-FT).[3]

---

[3]Hereafter whenever we refer to Head Start children we mean Head Start Planned Variation children.

For comparative purposes we have defined these groups quite strictly. A child is considered to have had Head Start only if he or she was in an HSPV program and his or her record indicates participation in Head Start for at least eight months. Also, children have been considered as having no preschool only if their data record indicates no preschool of any kind. While these stringent definitions restrict the size of our sample somewhat, they make comparisons less ambiguous.

Data for each of these treatment groups (NPS-NFT, NPS-FT, HS-NFT, and HS-FT) are given by site; cohort III in Tables A-1-A-4, and cohort IV in Tables A-5-A-8. Data are included in these tables only if they are avaialble for at least ten children in each treatment group in each site. As mentioned previously, data for this study are available for only seven sites in cohort III and twelve sites in cohort IV. Note, however, that HS-FT data are available in both cohort III and cohort IV in only six sites.

Table 1 summarizes the maximum sample sizes for which data are available for each treatment group, by cohort and site. This table makes clear why the task of estimating whether Follow Through extends the benefits of Head Start into early elementary grades is so difficult. Among the children for whom we have data, most Head Start graduates participated in Follow Through programs. In cohort III, for example, in only two of the seven HSPV-FT sites were there as many as ten children in the HS-NFT treatment groups. Moreover, in only one cohort III site are data available for at least ten children in all four treatment groups.

Table 1: Maximum Sample Sizes for Each of Four Treatment Groups by Site and Cohort*

Cohort & Site		NPS-NFT	NPS-FT	HS-NFT	HS-FT
Follow Through C III	204	67	34	-	72
	213	48	19	-	51
	316	110	137	21	45
	804	16	19	-	53
	902	47	-	-	38
	906	-	12	13	32
	910	-	14	-	-
Follow Through C IV	204	196	100	-	38
	209	123	35	-	28
	213	192	59	-	51
	309	125	169	-	62
	316	352	299	18	35
	511	302	24	-	60
	804	59	44	34	55
	808	-	38	-	45
	902	65	48	20	30
	906	66	51	20	24
	1203	137	14	-	65
	1204	179	69	-	27

*This table summarizes the maximum sample sizes by treatment group, site and cohort, of the data sets shown in Appendix A. Dashes indicate that data were available on fewer than 10 children in that treatment group for that site and cohort.

In cohort IV, available data appear to be more evenly distributed across all four treatment groups. In four of the CIV sites there are more than ten children in the HS-NFT treatment groups and in three sites there are more than ten children in all four treatment groups. Of the two cohorts, the fourth then offers a more promising data set in which to explore the effects of Follow Through in maintaining the benefits of Head Start. However, before pursuing this issue, let us describe the full HSPV-FT data set in more detail.

First of all, let us compare the present data with those from the same cohorts and sites included in the Smith (1973) and Weisberg (1974) analyses of HSPV data. These data, including sample size, mean age, precent black, percent white, mean mothers education, mean income, and spring and fall Head Start PSI scores, are given in Tables 2 and 3. For each site three sets of these data are given; first for the set of children whose data were analyzed in the HSPV evaluations; and then by HS-FT and HS-NFT groupings; that is, those HSPV children who turned up in either the FT or NFT data sets. As in previous cases we have not included data for groups in which sample size fell below ten.

In these tables several facets of the Head Start-Follow Through connection are evident. First of all, it seems clear that in most cases, the HSPV children who entered Follow Through are very much like the HSPV children as a whole. In both cohorts III and IV, however, the HSPV children who did <u>not</u> enter Follow Through but instead joined the comparison group NFT children, tended to be somewhat different from the original HSPV children.

In cohort III in sites 0316 and 0906, the ethnic composition (percent white) of the HS-NFT groups differs from that of the HSPV children by more than 15 percent. Curiously enough, the apparent differential selection of HSPV children into FT and NFT groups seems to work in different ways in these two sites; in site 316 the HS-NFT has a higher proportion black and lower proportion white than in the original HSPV sample in that site, whereas in site 0906 the proportion white in the HS-NFT group is three times higher in the HS-NFT than in the HSPV sample. Note also that there was an apparent differential selection in site 906 that was associated with entering-HS PSI scores--the children who went on into Follow Through averaged 5 points higher on the fall Head Start PSI than did the children in the same site who went into the comparison NFT group. Also, the tested children in the HS-NFT group seem to have gained more on the PSI during their year in Head Start than did the HS-FT children.

In cohort IV, as is shown in Table 3, a somewhat similar pattern is evident. Within most sites, HS-FT groups seem to be representative of the original HSPV samples, although there is considerable variation between sites on virtually every background variable. Again, however, note that in the four sites where there are more than ten children in the HS-NFT there appears to have been some sort of differential selection into FT and NFT. In site 804, for example, the proportion black in the HS-NFT sample is twice as high as in the original HSPV. This apparent differential selection also seems reflected in the PSI scores; in three out of the four sites with both HS-FT and HS-NFT groups, one of these traced groups

differs from the whole HSPV group on either fall or spring PSI score by at least one quarter standard deviation.

We shall return in a moment for a closer look at these HS-FT data, but first let us point out some other rather striking patterns evident in these data. Note, for example, that in this data set some of the background variables are almost entirely confounded by site and sponsor. For example, in cohort IV, sites 1203 and 1204, the two sites in this sample affiliated with the University of Pittsburgh, are 100 percent white, while the only site in this sample associated with Bank Street is 100 percent black. Also, several other variables, including fall Head Start PSI scores are badly confounded by site. In light of the extremely severe confounding of variables in this data set, we think it inappropriate to attempt any calculation of "sponsor effects" or direct model comparisons, and hence in the remainder of this paper refrain from doing so.

Follow Through Effects on Head Start

In this subsection we focus more closely on the HS-FT and NPS-NFT groups in those sites where substantial data are available for children in the HS-NFT groups; specifically four sites, 0316, 0804, 0902, and 0906.

Recall that in the last section, Tables 2 and 3 suggested some kind of differential selection of HSPV children into FT and non-FT. When we compare patterns across cohorts, we are given pause to consider what may actually be entailed in whatever differential selection may have been in effect. Consider site 0316 in both Table 2 and Table 3. In cohort III (Table 2), there is an indication of differential selection of black children into the

Table 2: Head Start Data for Original HSPV Sample, HS-FT and HS-NFT Groups, Cohort III.*

Site	Sample Group	Range of n**	Mean Age	% Black	% White	Mean Mom's Ed.	Mean Inc.	Head Start Mean PSI Scores (64-item) Fall	Spring
0204	HSPV	33	56.2	12.1	87.9	10.2	56.5	37.6	48.1
	HS-FT	72/27	55.1	6.9	80.6	10.8	61.4	37.4	48.1
	HS-NFT	<10	-	-	-	-	-	-	-
0213	HSPV	47	55.3	44.7	53.2	11.1	42.1	36.0	44.5
	HS-FT	51/19	55.3	27.5	64.7	10.9	39.2	37.5	44.3
	HS-NFT	<10	-	-	-	-	-	-	-
0316	HSPV	70	54.9	10.0	84.3	11.2	46.2	34.1	45.9
	HS-FT	45/20	53.4	13.3	71.1	11.0	44.2	31.5	45.1
	HS-NFT	21/13	54.0	28.9	66.7	11.2	40.2	33.1	41.1
0804	HSPV	63	55.4	41.3	58.7	9.7	34.3	32.1	45.1
	HS-FT	53/52	53.9	41.5	58.5	9.6	35.5	31.3	44.4
	HS-NFT	<10	-	-	-	-	-	-	-
0902	HSPV	58	53.0	75.9	24.1	10.1	33.1	23.7	40.7
	HS-FT	38/35	51.5	89.5	10.5	10.0	34.3	21.4	38.2
	HS-NFT	<10	-	-	-	-	-	-	-
0906	HSPV	39	56.1	0	15.4	9.4	35.4	40.5	42.5
	HS-FT	32/26	54.9	0	34.4	9.4	37.1	42.9	44.5
	HS-NFT	13/10	53.5	0	46.2	9.6	40.9	36.7	41.3
0910	HSPV	41	54.6	58.6	41.5	11.4	69.8	40.4	53.1
	HS-FT	<10	-	-	-	-	-	-	-
	HS-NFT	<10	-	-	-	-	-	-	-
Overall Standard Deviation								12.2	

*The three rows of data for each site represent overlapping subsets of a slightly larger set of data. The HSPV data come from Smith's (1973, pp. 59, 194) final analysis sample. The HS-FT and HS-NFT data come from the same set of data but are now part of our merged data set.

**Range of n refers to the size of the samples for which data are available. The larger n usually refers to age and ethnicity while the smaller n is usually for mother's education, income or PSI test scores.

***Overall standard deviation refers to overall individual standard deviation for all children tested in the fall 1970 HSPV sample. See Walker, Bane and Bryk, 1973, p. 338.

Table 3: Head Start Data for Original HSPV Sample, HS-FT and HS-NFT Groups, Cohort IV.*

Site	Sample Group	Range of n**	Mean Age	% Black	% White	Mean Mom's Ed.	Mean. Inc.	Head Start Mean PSI Scores (32-item) Fall	Spring
	HSPV	82/53	55.7	7.3	91.5	11.1	41.5	15.9	22.6
0204	HS-FT	38/24	54.9	7.9	89.5	10.9	37	15.4	21.8
	HS-NFT	<10	-	-	-	-	-	-	-
	HSPV	89/58	56.3	9.0	59.6	10.9	31.6	14.6	20.1
0209	HS-FT	28/22	55.9	3.6	53.6	10.9	32.5	14.8	21.3
	HS-NFT	<10	-	-	-	-	-	-	-
	HSPV	86/66	56.1	24.4	73.3	11.2	35.3	13.4	20.0
0213	HS-FT	51/47	54.8	23.5	74.5	11.1	34.8	13.1	19.7
	HS-NFT	<10	-	-	-	-	-	-	-
	HSPV	87/58	55.4	54.0	20.7	10.5	57.4	12.4	17.2
0309	HS-FT	62/36	53.8	51.6	21.0	10.4	59.2	11.3	15.3
	HS-NFT	<10	-	-	-	-	-	-	-
	HSPV	100/73	54.5	8.0	90.0	11.1	37.8	15.3	20.3
0316	HS-FT	35/32	53.5	17.1	74.3	10.2	35.8	13.9	19.0
	HS-NFT	18/16	54.4	0	100	11.1	36.4	14.8	20.4
	HSPV	102/71	51.8	100	0	10.6	33.1	9.5	13.3
0511	HS-FT	60/53	51.6	98.3	1.7	10.6	31.7	8.6	13.0
	HS-NFT	<10	-	-	-	-	-	-	-
	HSPV	64/52	56.0	35.9	64.1	9.9	39.1	14.3	21.7
0804	HS-FT	55/50	55.1	40	60	9.6	37.7	13.1	21.2
	HS-NFT	34/33	53.9	73.5	26.5	9.4	29.5	11.6	16.2
	HSPV	71/49	52.4	70.4	29.6	10.6	46.9	14.0	15.8
0808	HS-FT	45/43	52.4	80	20	10.1	44.5	13.1	15.0
	HS-NF%	<10	-	-	-	-	-	-	-
	HSPV	85/62	53.3	75.3	24.7	10.4	31.7	11.5	15.0
0902	HS-FT	30/27	53.3	96.7	3.3	10.1	29.4	10.2	14.3
	HS-NFT	20/18	51.5	60.0	40	9.9	34.3	11.9	14.6
	HSPV	49/40	56.6	0	18.4	10.2	36.7	13.1	18.1
0906	HS-FT	24/21	55.3	0	33.3	10.4	37.1	13.5	18.8
	HS-NFT	20/19	54.9	0	20	8.8	47.3	13.3	16.5
	HSPV	91/75	52.0	0	100	11.0	41.9	11.3	17.8
1203	HS-FT	65/57	51.2	0	100	11.0	41.3	11.4	18.3
	HS-NFT	<10	-	-	-	-	-	-	-
	HSPV	48/41	55.9	0	100	11.6	47.9	18.0	23.4
1204	HS-FT	27/22	57.1	0	100	11.6	52.2	19.6	24.7
	HS-NFT	<10	-	-	-	-	-	-	-
Overall Standard Deviation								6.2	

*As in Table 2, the three rows of data for each site represent overlapping subsets of a somewhat larger set of data. The HSPV data are from Weisberg's (1973, pp. 44, 51) HSPV evaluation report. The HS-FT and HS-NFT data come from our current HS-FT study. The range of sample sizes represent the largest and smallest number of children from whom background data in each sample group were drawn. In most cases the larger n applies to demographic data and the smaller to test data.

**Range of n refers to the size of the samples for which data are available. The larger n usually refers to age and ethnicity while the smaller n is usually for mother's education, income or PSI test scores.

***Overall standard deviation refers to overall individual standard deviation for all children tested in the fall 1970 HSPV sample. See Walker, Bane and Bryk, 1973, p. 338.

HS-NFT group, but in cohort IV, there are no black children in the HS-NFT sample in that site. In fact, in cohort IV white children seem to have been differentially selected into the HS-NFT group. This suggests that whatever the impact of this pattern on the data set available to us, it may have been caused more by year-to-year fluctuations of sampling than by any long-term institutional selection mechanism. Also, some sort of sampling variation is suggested in the range of n's available for different types of data.

The sampling problem became especially apparent to us when we began examining the Follow Through data available for the HS-NFT sample groups. Although some sort of Follow Through data are available for more than 30 HS-NFT children in cohort III (in sites 0316 and 0906), as soon as we begin to restrict attention to any sort of traced sample (with test data at more than two time points), the cohort III HS-NFT sample quickly shrinks to less than ten children per site--a size which makes any sort of meaningful comparison of HS-FT and HS-NFT children over time virtually impossible.

The situation in cohort IV is somewhat better--or at least marginally so. When restricted to traced data for which we have HS, kindergarten, first and second grade data, we still have sample sizes of at least ten in three of the CIV sites--804, 902 and 906. In the 316 site, selection for traced data reduced the HS-NFT sample size to only five.

The data for the traced HS-FT and HS-NFT samples for the 0804, 0902 and the 0906 sites are given in Table 4. Several points should be noted about this traced set of data. Comparing Table 4 with Table 3, it can be seen that restricting ourselves to the traced set of data reduces sample sizes in each of the three sites. In terms of available background data,

however, the traced samples in each of the three sites show composition closely resembling that shown in Table 3. This includes the apparent differential selection associated with ethnicity. In site 804, black Head Start graduates seem to have been differentially selected into the NFT comparison group while in site 0902 black Head Start graduates appear to have been differentially selected into Follow Through. In site 906 a majority of the traced sample is neither black nor white--rather it is Hispanic-American.

These three sets of traced data offer us then only an extremely limited opportunity to estimate the degree to which Follow Through maintains the benefits of Head Start. Not only is there evidence of some sort of differential selection of Head Start graduates into FT or NFT in each of three sites, but also the traced samples in the three sites are themselves markedly different in terms of ethnicity. Also, these three sites are in different regions of the U.S. Site 804 is in Portageville, Missouri; 902 in Oskaloosa County, Florida; and 906 in Greeley, Colorado.

Nevertheless, despite these severe limitations in terms of comparability, we still think it useful to offer some very tentative conclusions about the effectiveness of Follow Through in "maintaining the benefits of Head Start". Consider the right-hand half of Table 4. MAT scores for each of the traced sample groups are given for each site, at spring kindergarten, spring first grade and spring second grade (recall that CIV was not tested at spring third grade). Total Reading scores for each sample group are given in terms of standard scores and percentile equivalents provided by the publisher. These data suggest that Follow Through has had

Table 4: Cohort IV Traced Head Start Children: HS-FT versus HS-NFT. READING

Site	Sample Group	n	Mean Age	% Black	% White	Mean Mom's Educ.	Mean Inc.	Head Start Mean PSI Scores Fall	Head Start Mean PSI Scores Spring	MAT Reading Total Scores Spring-K SS	Spring-K %	Spring-1 SS	Spring-1 %	Spring-2 SS	Spring-2 %
804															
	HS-FT	37	56.4	50	50	9.4	37	14.6	22.1	30	84	46	52	54	60
	HS-NFT	27	55.2	75	25	10.3	32	12.7	17.5	23	48	42	54	53	50
902															
	HS-FT	25	54.1	96	4	10.3	29	8.8	13.8	22	39	29	17	42	11
	HS-NFT	15	52.9	64	36	9.9	30	13.0	16.1	25	48	37	37	47	22
906															
	HS-FT	11	55.9	0	30	9.6	43	12.8	18.2	25	50	32	23	45	16
	HS-NFT	17	56.5	0	13	8.6	50	13.5	16.3	23	42	34	31	47	24

some impact in maintaining children's MAT scores through the end of kindergarten--especially in site 0804 and to a lesser extent in site 0906. There is, however, some indication that the scores of the traced HS-FT students do decline, relative to both the respective HS-NFT comparison groups and publishers' percentile norms after the end of kindergarten. In site 804 this trend does not seem to continue beyond the end of first grade, but there does seem to be further relative decline in the 0902 and 0906 site traced groups.

It is important to reiterate, however, that the patterns revealed in these data must be viewed with considerable reservation. First of all, this is an extremely small and unique data set--from only three out of more than 100 Follow Through projects, from only a small number of children in each of these sites, and on only one outcome measure. The second problem is that the HS-FT and HS-NFT comparisons are suspect because of ways in which background characteristics of Head Start graduates are confounded in the HS-FT and HS-NFT groups.

In the next section, we return to the larger HSPV-FT data set to explore the Head Start confounding issue more fully.

III. Confounding Problems in Estimating Average Follow Through Effects

The flow of children into, through, and out of Head Start and Follow Through constitutes a vast and complex process. Children were selected for Head Start on the basis of general criteria applicable nationally, but local circumstances determined the specific make-up of program groups. Thus groups of Head Start children in different places vary widely on numerous dimensions. In Follow Through, too, the likelihood of participation depends on children's characteristics and local circumstances. Moreover, Head Start experience is one of the factors taken into account in the selection process.

Children in Follow Through, then, represent a special sample resulting from a complex and unique selection process. To measure the effectiveness of Follow Through on any dimension of interest requires a valid standard of comparison, so that actual performance can be compared with that which this same group would have achieved without Follow Through.

In the major Follow Through evaluations, the performance of comparison groups of NFT children has been used to develop such a standard. These NFT groups were also formed by a complex process and one that probably reflected a different balance of forces from that at work in the Follow Through selection process.

In the national Follow Through evaluation, statistical adjustments based on measured variables have generally been used to compensate for the differences between Follow Through and Non-Follow Through comparison children. The adequacy of these sorts of adjustments has long been subject

to scrutiny, and recently several analysts have pointed out that such adjustments may entail substantial biases unless carried out with knowledge of the processes of group selection (Bryk and Weisberg, 1977; Cronbach, Rogosa, Floden and Price, 1976; Magidson, 1977). Without such knowledge, the effects of treatments will be inextricably confounded with selection biases.

In this report, we hope to clarify how these general problems have operated in evaluations of Follow Through, and draw out some implications for interpreting results of the national evaluation. In the remainder of this section we set the stage for our data analyses. First we discuss the confounding problem as manifested in the national Follow Through evaluation and then develop a conceptual framework which is helpful in sorting out the key issues.

Confounding in the National Evaluation

The mechanism by which children were assigned to Follow Through or a Non-Follow Through comparison was far from random. Follow Through and Non-Follow Through children differ in terms of many variables in different ways across sites. Typically, the Non-Follow Through group, on the basis of their background characteristics--ethnicity, SES, and test scores--could be expected to perform considerably higher on achievement tests at third grade than Follow Through children. Also, the Follow Through group typically contains a higher proportion of Head Start graduates (see for example, Stebbins, et al., 1975).

Analysts of the Follow Through data at Abt have always recognized that the mean outcomes could not be compared without some compensation for the

initial differences on background characteristics. Throughout their various studies, they have employed a variety of statistical adjustment strategies. The main approach has been analysis of covariance (ANCOVA).

ANCOVA is based on two major assumptions. First, it is assumed that the expected outcome difference between groups in the absence of an intervention can be represented as a linear function of mean differences on a set of measurable pre-treatment variables (covariates). Second, it is assumed that the coefficients in this linear function can be estimated from the data. The usual approach is to calculate the pooled within-group regression coefficients, and to use these in performing the adjustment.

The adequacy of such statistical adjustments, and ANCOVA in particular, has been fiercely debated in recent years (c.f. Campbell and Erlebacher, 1970 and reply by Evans and Schiller, 1970). The most serious objections center around the issue of whether the covariates used contain enough information on growth potential to allow a reasonable adjustment. That is, have all variables with a causal influence on outcomes been included in the covariate set? This is a very difficult question to answer definitively. The adequacy of covariates used becomes in part a matter of judgment.

Since entry to kindergarten represented the beginning of the Follow Through program as far as Abt was concerned, they selected covariates describing the child at this time point. The exact covariate set varied across their many analyses, but almost always included demographic variables and the Wide Range Achievement Test (WRAT), administered at children's entry into kindergarten as an index of pre-treatment level of academic performance.

A key question in using WRAT scores in this way is whether they represent base-line measures of status which adequately predict third grade outcome with no Follow Through intervention. For children with no Head Start, this may be a reasonable assumption. For children with Head Start or some other comparable preschool program, however, the WRAT score may represent in part an outcome of this program. Thus its relation to third grade outcome might be quite different than for children without preschool experience. In particular, if we accept some form of the fade-out hypothesis, we might expect the WRAT to over-predict third grade performance. The resulting bias would tend to make sites with more Head Start children look bad, and in general to lower the estimated effect of Follow Through (since Follow Through groups generally have a high proportion of children with Head Start experience, and comparison groups a relatively low proportion).

By now the reader may be quite confused. We have pointed to several conceptual and technical issues which in combination raise doubts about the meaningfulness of the national Follow Through evaluation results. Of special importance here is whether these various considerations seriously affect the general credibility of the national Follow Through evaluation. To determine this, we must unravel the different factors mentioned, and the ways they interact. In the next subsection, we present a conceptual framework which will be helpful for this purpose.

Conceptual Framework

Suppose one were trying to design a study of Follow Through which took explicit account of Head Start. In order to explore the relationships of Follow Through and Head Start effects completely, it would be necessary to define four essentially different possible treatments which a child might receive between the ages of 4 and 9. These are:

(1) HS-FT	Head Start and Follow Through
(2) HS-NFT	Head Start and no Follow Through
(3) NPS-FT	No Preschool and Follow Through
(4) NPS-NFT	No Preschool and no Follow Through

Of course both HS and NPS are shorthand notations for a variety of programs, varying in terms of curriculum emphasis, goals, amount of time per day, and length of the program. For simplicity in this discussion, we will ignore this variation, and treat HS and NPS as well-specified entities.

Also, the discussion which follows will not deal with the issue of the appropriate level for estimating effects. It is an important theoretical issue whether effects vary across sponsors, or sites. If, for example, a model's effectiveness depends strongly on where it is implemented, an overall sponsor effect is not meaningful. However, in order to focus on the HS-confounding issue, we assume here that a theoretically appropriate level for the estimation of overall effects has been determined. Our argument could, in principle, be applied to sites, sponsors, or other units.

Now let τ_{HS} represent the average difference between what a child would score at the end of third grade with Head Start and without Head Start. Similarly, let τ_{FT} represent the effect of Follow Through alone on spring third grade scores. Finally, let τ_{HS+FT} be the cumulative effect of Head Start and Follow Through on spring third grade scores.

Note that we are not assuming that $\tau_{HS+FT} = \tau_{HS} + \tau_{FT}$, although this additive model represents one possibility. The relationship between τ_{HS}, τ_{FT}, and τ_{HS+FT} is an interesting theoretical issue which an evaluation might address. The various vague conceptions about Head Start and Follow Through considered in the first section can be made more precise by stating them in terms of these parameters. For example, a strong form of fade-out might be represented by

$$\tau_{HS} = 0$$

A take-off model on the other hand might be represented by specifying a particular positive value for τ_{HS}.

To estimate τ_{HS}, τ_{FT}, and τ_{HS+FT}, we would ideally like to have four comparison groups: one with each of the four types of experience mentioned above. Moreover, these groups should be equivalent with respect to growth potential on the dimension being measured, as of the age of preschool entry. This could best be accomplished by random assignment.

With this ideal experiment, a comparison of groups (1) and (4) would estimate τ_{HS+FT}. A comparison of (2) and (4) would estimate τ_{HS}, and (3) and (4) would estimate τ_{FT}.

Now let us consider what has actually happened in the major Follow Through evaluation. In general, an overall Follow Through effect has been estimated by comparing whatever group happened to receive Follow Through in a given site with the Non-Follow Through group (for the moment we ignore selection biases).

Let

p = proportion of FT children with HS experience

$1-p$ = proportion of FT children with no PS

q = proportion of NFT children with HS experience

$1-q$ = proportion of NFT children with no PS

Note that for simplicity we are continuing to ignore children in PS programs other than HS.

Suppose that we are interested in an overall average FT effect. The simplest estimate of such an overall effect would be a difference of mean outcomes for children in FT and NFT. Under our model, this difference of means would be an estimator of

$$\varepsilon = p\,\tau_{HS+FT} + (1-p)\,\tau_{FT} - q\,\tau_{HS}$$

That is, in terms of the three distinct treatment effects we have defined, the overall comparison is somewhat ambiguous. The interpretation of this quantity will vary, depending on what assumptions we make about the τ's. For example, suppose the effects of HS and FT are simply additive. Then

$$\tau_{HS+FT} = \tau_{HS} + \tau_{FT} \quad \text{in which case} \quad \varepsilon = \tau_{FT} + (p-q)\,\tau_{HS}$$

That is, we are estimating the effect of FT alone only if the proportions of HS graduates in FT and NFT are equal.

Suppose, however, that HS and FT effects are not additive, and that a strong fade-out takes place. If under the strong fade-out assumption

$$\tau_{HS} = 0, \text{ then } \varepsilon = p\,\tau_{HS+FT} + (1-p)\,\tau_{FT}$$

Now ε is estimating a weighted average of the effect of FT alone and of HS and FT combined. Of course, if both of the above conditions hold,

$$\tau_{HS} = 0, \text{ and}$$

$$\tau_{HS+FT} = \tau_{HS} + \tau_{FT} = \tau_{FT}$$

Then $\varepsilon = \tau_{FT}$.

Thus under strong fade-out and additivity, estimation is straightforward, and the proportions in different program groups are irrelevant.

So far we have ignored background differences among the four program groups. Now, however, let us suppose that children in the HS-FT group have characteristics which on the average would produce a higher (or lower) MAT third grade outcome than that of the NPS-NFT group, even if children were not given any special treatment. Let us denote by S_{HS} the average difference on the MAT attributable to these group differences. Similarly let us define S_{FT} and S_{HS+FT}.

Now children in HS-FT can be expected to score $\tau_{HS+FT} + S_{HS+FT}$ higher, on the average, than NPS-NFT children, assuming that treatment and selection effects are additive. NPS-FT children will score $\tau_{FT} + S_{FT}$ higher, and HS-NFT children $\tau_{HS} + S_{HS}$ higher. Thus a simple comparison of FT and NFT means would estimate:

$$\varepsilon = p\,(\tau_{HS+FT} + S_{HS+FT}) + 1\text{-}p\,(\tau_{FT} + S_{FT}) - q\,(\tau_{HS} + S_{HS})$$

The overall effect of these selection biases depends on assumptions about the S's. For example, it may be that conditional on HS or NPS, biases in FT selection are minimal, i.e., suppose

$$S_{HS+FT} = S_{HS} = S < 0$$

$$S_{FT} = 0$$

Then

$$\varepsilon = p\,\tau_{HS+FT} + (1\text{-}p)\,\tau_{FT} - q\,\tau_{HS} + S_{HS}(p\text{-}q)$$

That is, our estimate will have a bias whose magnitude and sign are a function of the difference between HS proportions in the FT and NFT groups. As long as a higher proportion of FT children than NFT children have had HS, as is in reality usually the case, the bias will be negative.

Now let us consider the statistical adjustment. This adjustment will be a complex function of groups means on the entering kindergarten WRAT and other background variables. Let us represent the net effect of

this complex function as γ. Ideally γ should just compensate for selection biases resulting from the assignment process. With adjustment, then

$$\varepsilon = p\ (\tau_{HS+FT} + S_{HS+FT}) + 1\text{-}p\ (\tau_{FT} + S_{FT}) - q\ (\tau_{HS} + S_{HS}) - \gamma$$

In most of Abt's major analyses, the WRAT is the major covariate. It displays the strongest statistical relationship with third grade outcomes. The meaning of the WRAT as a predictor is clouded by the fact that if Head Start has short-term effects, it may influence children's performance on the WRAT at the time they begin Follow Through.

If we ignore the other covariates besides the WRAT, the adjustment γ is of the form

$$\gamma = \beta(\bar{X}_{FT} - \bar{X}_{NFT})$$

where

$\bar{X}_{FT}$ = WRAT mean for FT children

$\bar{X}_{NFT}$ = WRAT mean for NFT children

and β is computed from the data using a complex statistical procedure. Typically, β is based on the relationship between pre- and post-tests within the two groups.

Now we said above that for children from HS, WRAT scores may be inflated. That is, if we knew the scores they would have obtained without HS, we could use these in the adjustment technique. Let us represent these hypothetical "HS-free" WRAT's by W's. Moreover, assume (without loss

of generality) that the scores are scaled so that the mean is zero for the NPS-NFT group. Finally, let us represent by α the increment to the WRAT score provided by HS.

It follows from these definitions that

$$\bar{X}_{FT} = p\ (W_{HS+FT} + \alpha) + (1-p)\ W_{FT}$$

$$\bar{X}_{NFT} = q\ (W_{HS} + \alpha)$$

Thus

$$\gamma = [p\ (W_{HS+FT} + \alpha) + (1-p)\ W_{FT} - q(W_{HS} + \alpha)]$$

Combining this with our previous expression for ε yields finally

$$\varepsilon = p\ (\tau_{HS+FT} + S_{HS+FT} - \beta W_{HS+FT} - \beta\alpha) + 1-p\ (\tau_{FT} + S_{FT} + \beta W_{FT}) -$$

$$q\ [\tau_{HS} + S_{HS} - \beta W_{HS} - \beta\alpha)$$

This formula shows that even under the simple assumption that the influences of various factors are additive, the adjusted differences between post-test scores for Follow Through children and post-test scores for Non-Follow Through children is difficult to interpret. It also helps to explain exactly why the relationship between Head Start and Follow Through must be considered. For an evaluation of third grade effects, the objects of real interest are the "true" effects, τ_{HS+FT}, τ_{HS}, and τ_{FT}. While a randomized experiment would allow us to estimate these quantities directly, the quasi-experiment actually implemented involved three "extraneous" factors which influenced the results. These are:

a) The proportions of children experiencing each of the four possible combinations of preschool and early elementary experience.

b) Differences among the streams of children flowing into each of the preschool-early elementary groups.

c) The influence of Head Start on entering kindergarten performance.

The precise effect of these forces in combination is difficult to guage. The three forces are not independent. They are all reflections of the operation of a complex system, and it may be extremely difficult to isolate them. To interpret the results of the national Follow Through evaluation it would be useful to describe as precisely as we can the way this system has operated. In the next section, we analyze the HSPV-FT data in this light.

IV. Analysis of the Data

In the previous section, we discussed the ways in which the relationship between Head Start and Follow Through confused the issue of estimating overall Follow Through effects. We showed that three factors related to Head Start may influence the meaning of the kind of adjusted mean differences calculated in the national Follow Through evaluation:

a) The proportions of children experiencing each of the four possible continuations of preschool and early elementary experience.

b) Differences among the streams of children flowing into each of the preschool-elementary groups.

c) The influence of Head Start on entering kindergarten performance.

In this section our goal is to explore the kinds of distortions these three factors may introduce. Precise quantitative estimates of biases would require a very complex mathematical model and much richer data. Our goal here is more modest. We will comb the available data for clues as to the direction, if not magnitude, of the three factors.

Our approach is exploratory, and deliberately simple. We want to organize the data in such a way that the reader can easily judge the evidence and reach an independent conclusion. Since there are no clear answers, we want to let the numbers speak for themselves as much as possible.

The main analyses conducted by Abt used the entering kindergarten WRAT scores as the principal covariate. The mean WRAT for a particular Follow Through or Non-Follow Through group was in effect taken as a baseline measure of pre-program performance. As we have seen, however, this WRAT mean could have been influenced by all three of the key factors we have been discussing: proportion of children with Head Start; background differences; and effect of Head Start. Thus a good place to start seemed to be an attempt to elucidate the meaning of these scores.

Since Cohort IV had no fall kindergarten testing, we begin by exploring the Cohort III data. Ideally, we would have information on performance both prior to and at kindergarten, in each of the four treatment groups. However, as noted in section II, we had only a negligible number of children in the HS-NFT group. Furthermore, we had no data on the performance of the NPS groups prior to FT. Thus, the possibilities for analysis were rather limited.

Given these limitations, the simplest way of learning something about the meaning of WRAT fall kindergarten (FK) scores, was to hold them constant, and to see what values of other variables seemed to cluster with particular values. We began by dividing each group on the basis of FK WRAT into several blocks. The number of blocks was a compromise between homogeneity and sample size. We wanted score intervals to be small, so that scores were nearly the same, but needed substantial sample sizes. For each block, we computed the mean values of several variables characterizing individuals at FK or before. Most important as suggestive indices were FK Peabody Pictures Vocabulary Test (PPVT), FK Preschool Inventory (PSI), and ethnicity.

The PPVT and PSI were found in Huron's previous studies of HSPV (Smith, 1973; Weisberg, 1974) to be useful indices of academic readiness. A detailed discussion of these measures is provided by Walker et al. (1973). In general, both Smith and Weisberg found evidence of gains in PSI scores resulting from HS. The PPVT was not used in Cohort III, and evidence of HS effect on it in Cohort IV was equivocal.

Ethnicity (actually percent white) was used as a rough proxy for a variety of background variables relating to school problems, since it is highly reliable and easy to obtain. Information on SES variables like income and mother's education were examined in early exploratory analysis, but problems of reliability, availability, and comparability between HS and FT data bases led us to abandon these in the final presentations.

In addition, we wanted to trace growth for those who started with a given WRAT score. Thus, we also calculated MAT standard scores at SK, S1, S2, and S3. At this point, we realized that to trace growth meaningfully, we ought to be looking at the same children at different time points. We decided to restrict the analysis to those children for whom we had MAT scores at all four points, as well as entry WRAT data.

In addition to comparisons across treatment groups, we became interested in the patterns of growth over time within groups. We decided to use the percentiles based on the Harcourt Brace standardization as a rough standard. Do children who start out at a certain level tend to maintain a roughly constant percentile value? If not, can we interpret the observed variation?

Examination of the data arranged in this way generated several interesting findings. We therefore decided to apply the same approach to the Cohort IV data. As described in section II, Cohort IV data contain no information from FK or from S3, but do contain a non-negligible sample of HS-NFT children. Thus a comparison of HS-FT and HS-NFT was feasible. So instead of blocking on FK WRAT as in CIII, we blocked the CIV scores on FHS PSI. For each CIV block, mean scores were calculated for PSI SHS, PPV FHS, PPV SHS, ethnicity and MAT standard scores for SK, S1, and S2. Since there were no preschool data available for the NPS-FT and NPS-NFT groups, the CIV FHS-blocked data are presented for the HS-FT and HS-NFT groups only.

In the next subsection, we present tables of the data organized as described above. We make a number of observations which bear on the forces influencing group selection processes and program effects. These are then pulled together in a concluding section summarizing what we think these results mean for the national evaluation.

Discussion of Tables

Let us consider first the results for Cohort III. Altogether, we present six tables for Cohort III; for three program groups (NPS-NFT, NPS-FT, and HS-FT) and two domains of competence (the MAT Reading and Math scores). The tables are identical in format, except for the fall kindergarten score intervals used to form blocks, which differ for reading and math.

As an example, consider Table 5, which presents results on reading for the NPS-NFT group. Each row with an $\bar{X}$ at the left contains means for different variables for children with the range of WRAT scores indicated in the left hand column. At the far right is the sample size of each block.

For example, the first row represents children with scores from 1 to 9 on the WRAT at entry (fall) to kindergarten. Zero scores were eliminated, because we thought their meaning might be ambiguous. This group of 26 children was 77% white (21 of 26), had a mean PPVT of 42.5, and a mean PSI of 15.4. MAT Standard scores were 21.4, 35.6, 48.2, and 54.3 at the four test points.

In the lower line of each row, following the p is the value of the normative rank corresponding to the mean score according to publishers'

norms (Durost, et al., 1971). For example, a score of 21.4 on the MAT reading at spring of kindergarten would place an individual at the 37th percentile of the norming sample.

Let us begin by comparing the NPS-NFT and NPS-FT groups (Tables 5 and 6). Looking at the first row of each, we see that scores on variables available at FK (PPVT, PSI % white) are quite similar for these groups. This similarity seems to hold true for each of the other blocks as well.

On the PPVT, the largest discrepancy is 4.5 points (for the 10-12 block), and five of the six blocks are within 2.5 points. On the PSI, the discrepancy is less than 2 points for all six blocks. Per cent white is high for all blocks, and the maximum discrepancy is 6 percentage points.

It seems then that the kind of children getting a particular WRAT score in each of the two groups is similar. In other words, the meaning of the WRAT as a predictor of future performance may be similar for NPS-NFT and NPS-FT children. This situation suggests that for NPS children it may be reasonable to use the WRAT as a covariate in an analysis of covariance.

Looking at Table 7, the HS-FT group, we have quite a different picture. Comparing comparable blocks on the PPVT, we find the HS-FT children consistently lower than for either NPS group. Discrepancies range from a low 2.4 (HS-FT versus NPS-FT 1-9 block) to a high of 7.3 (HS-FT versus NPS-FT 10-12 block). On the PSI there is much greater similarity, with the exception of the highest (24-39) performers. The percent white is much lower in HS-FT than in NPS-FT or NPS-NFT.

Table 5: Cohort III Blocked NPS-NFT Group, Traced Reading Scores

WRAT Reading FK		PPV FK	PSI FK	MAT SK	MAT S1	MAT S2	MAT S3	Income*	% White	n
1-9	$\bar{X}$	42.5	15.4	21.4	35.6	48.2	54.3	7800	77	26
	P			37	34	29	29			
10-12	$\bar{X}$	48.7	17.5	24.3	38.0	50.2	57.3	8700	90	30
	P			47	42	39	37			
13-15	$\bar{X}$	54.2	20.4	24.1	43.6	55.7	62.7	9300	88	16
	P			46	68	65	57			
16-19	$\bar{X}$	52.8	20.5	26.3	42.1	52.5	61.2	8300	100	11
	P			61	60	50	51			
20-23	$\bar{X}$	58.2	22.6	27.5	49.1	62.4	72.8	10,000	100	11
	P			73	88	83	82			
24-39	$\bar{X}$	57.1	23.0	32.1	56.1	64.0	73.8	9600	100	19
	P			94	95	86	84			
										113

*This figure is based on total sample for whom WRAT FK and income data were available, not traced group only. It is only approximate since data were grouped into intervals.

Table 6: Cohort III Blocked NPS-FT Group, Traced Reading Scores

WRAT Reading FK		PPV FK	PSI FK	MAT SK	MAT S1	MAT S2	MAT S3	Income*	% White	n
1-9	$\bar{X}$	41.8	13.4	23.1	34.9	44.2	55.3	8200	73	15
	P			40	32	15	31			
10-12	$\bar{X}$	53.2	17.4	25.5	36.7	48.3	54.7	8700	90	31
	P			54	39	30	29			
13-15	$\bar{X}$	53.0	18.8	28.9	42.4	50.5	60.7	8300	86	21
	P			85	61	41	49			
16-19	$\bar{X}$	55.1	20.9	26.9	41.8	54.5	65.2	9000	100	12
	P			67	60	59	54			
20-23	$\bar{X}$	55.7	21.6	28.9	42.9	53.0	61.3	9200	94	17
	P			85	66	52	51			
24-39	$\bar{X}$	57.9	22.5	34.2	53.0	64.8	74.0	9800	96	24
	P			97	92	87	84			
										120

*This figure is based on total sample for whom WRAT FK and income data were available, not traced group only. It is only approximate since data were grouped into intervals.

One possible interpretation would be that WRAT and PSI scores have been raised by children's HS experience. A given WRAT score may be associated with generally more disadvantaged children in HS-FT than is the same WRAT score in either of the NPS groups. Thus the WRAT alone would probably be inadequate as a covariate in comparisons between HS-FT and NPS-NFT children. To use the WRAT for adjustment, we would need more information on what its real meaning is in the context of possible HS effects and the group selection process.

In comparing the three groups, another factor stands out. Notice that for the NPS-NFT group 17% (19 of 113) are in the highest category. For the NPS-FT group 20% (24 of 120) are in the highest category. For the HS-FT group, however, only 8% (9 of 115) are in the highest category. Moreover, the mean PPVT and PSI scores for these 9 are only 51.8 and 20.0. Corresponding values for the NPS groups are considerably higher. All this suggests that selection factors have resulted in a subgroup of NPS children essentially out of range of the HS-FT sample. For such children, no statistical adjustment is likely to be adequate.

Now let us look at the MAT outcomes. Comparing Tables 5 and 6, we note that at SK, the NPS-FT group performs higher than the NPS-NFT group in every block. By S1 the situation has completely reversed; the NPS-NFT children are higher. Since the groups seemed quite similar at fall kindergarten, this pattern strongly suggests that FT has boosted scores during kindergarten, but that the gains have not shown up later. At S3, the results are mixed. There is no clearcut advantage or disadvantage of Follow

Table 7: Cohort III Blocked HS-FT Group, Traced Reading Scores

WRAT Reading FK		PPV FK	PSI FK	MAT SK	MAT S1	MAT S2	MAT S3	Income*	% White	n
1-9	$\bar{X}$	39.4	15.3	22.1	27.9	41.7	49.4	5300	24	21
	P			38	12	11	15			
10-12	$\bar{X}$	45.9	16.3	24.6	34.6	46.7	52.3	6500	53	32
	P			48	31	22	23			
13-15	$\bar{X}$	48.8	19.9	26.3	36.1	48.7	55.2	6900	43	21
	P			61	36	32	21			
16-19	$\bar{X}$	49.1	20.6	27.7	37.3	50.8	58.6	6900	69	16
	P			75	40	43	41			
20-23	$\bar{X}$	53.7	21.6	27.7	42.4	54.3	61.3	7300	75	16
	P			75	63	58	51			
24-39	$\bar{X}$	51.8	20.0	29.9	46.8	54.9	63.6	6900	67	9
	P			90	82	62	59			
										115

*This figure is based on total sample for whom WRAT FK and income data were available, not traced group only. It is only approximate since data were grouped into intervals.

Through. Thus Follow Through seems to be doing something, but whatever it does is not evident in third grade outcomes.

It is also worth noting that in percentile terms, the NPS-NFT group remains fairly constant over time. This seems consistent with what we would expect in the absence of a special program. The NPS-FT group, on the other hand, starts out higher than the NPS-NFT group at SK, shows a drop at S1, and another drop at S2, but an apparent leveling off between S2 and S3. These patterns again suggest an early FT effect which is no longer evident on third grade tests.

Now let us look at the Head Start-Follow Through group. Although these children tend to be more disadvantaged than the NPS-NFT groups, their MAT SK scores are, with one exception, consistently higher! In fact, the median score for HS-FT children at SK overall is well above the 50th percentile! It is hard to explain this as anything but an effect of the Head Start and Follow Through experiences these children have had. By S3, however, they have dropped to much lower levels. Again we see an early boost followed by a drop-off. Also as with the NPS-FT group there is an apparent leveling off between S2 and S3. The key question is whether the HS-FT children level off above or below where they would have been without Follow Through. Unfortunately, we cannot answer this question since we do not have an appropriate HS-NFT control group.

We note also further evidence that the NPS groups contain children of much greater growth potential than any of the Head Start-Follow Through children. In both NPS-FT and NPS-NFT, the groups scoring above 23 on the WRAT tend to stay around the 90th percentile! These are outstanding

students, not comparable to FT children who score at least a full quintile lower in second and third grade.

Tables 8, 9, and 10 present comparable data for math. The results are quite similar to those for reading, and it would be redundant to discuss them in detail. However, the reader is invited to form his/her own interpretation of the information presented. We shall proceed to discuss the Cohort IV analysis.

For the NPS children in Cohort IV we have neither preschool data nor FK data. Thus there is not much that can be usefully analyzed for NPS-FT and NPS-NFT groups. On the other hand, we do have a small but non-negligible sample of HS-NFT children. Thus we can explore in a limited way the performance of FT children who have had HS, by comparing the HS-FT and HS-NFT groups.

Tables 11 and 12 present reading results for HS-NFT and HS-FT respectively. Tables 13 and 14 present math results. Let us consider the reading data first. The format of the tables is similar to those already discussed. Blocking is now on the basis of the PSI (32-item version) taken at HS entry. Other measures are the fall HS PPV, the spring HS PSI and PPV, and the MAT standard scores at SK, S1, and S2.

Given the small samples in the HS-NFT blocks, it is unwise to read too much significance into the exact values. In general, it appears that based on % white, PPV FHS, PPV SHS, and PSI HS, the NS-NFT and HS-FT groups are very different once we control for entry PSI. FT children are possibly slightly ahead. Specific blocks do show differences, but they are difficult to interpret. For example, among children scoring low in the FHS PSI

(0-6 and 7-9) those going into NFT seemed to have performed better in HS than those going into FT. For high-scorers, the situation is reversed. Is this a meaningful characteristics of the selection process, some kind of program effect, or just random variation?

Now let us look at MAT scores. Note that HS-FT children are consistently above the corresponding HS-NFT children at SK, in some cases quite substantially. Even the apparently very disadvantaged HS-FT lowest scores (0-6 on PSI) slightly outperform their counterparts in NFT.

By S2, however, the groups are very comparable. The FT group tends to be slightly ahead, just as it was at the outset. This suggests a negligible effect of Follow Through on second grade outcomes. The apparent effect of Follow Through was to raise scores at SK, but not at S2.

Note also the percentile values. For both groups the MAT SK scores seem inflated, and particularly so for the HS-FT group. Again the median SK score is well above the 50th percentile! As all HS children are presumably disadvantaged, this is quite remarkable. There seems to be additional evidence here that both HS and FT raise levels on whatever skills the MAT measures at SK.

Results for math are similar in general. Again the FT group seems higher at SK and drops back. Specific results vary considerably across blocks, however. Some show FT doing better than one might expect on the basis of background and HS scores (e.g., 16-18 block); others show FT doing worse (e.g., 10-12 block). Samples are too small to interpret such patterns with any confidence.

Table 8: Cohort III Blocked NPS-NFT, Traced Math Scores

WRAT Math FK		PPV FK	PSI FK	MAT SK	MAT S1	MAT S2	MAT S3	Income*	% White	n
1-4	$\bar{X}$	41.4	13.3	16.9	35.9	49.7	63.0	8400	86	7
	P			20	32	16	26			
5-7	$\bar{X}$	44.5	15.5	19.1	36.3	47.6	61.5	7600	87	15
	P			26	33	12	22			
8-10	$\bar{X}$	49.8	18.1	23.0	36.7	47.7	61.6	9700	87	23
	P			52	35	12	22			
11-13	$\bar{X}$	50.5	19.4	25.1	42.2	56.3	67.9	8200	100	18
	P			62	50	46	40			
14-16	$\bar{X}$	55.0	21.4	28.2	43.4	60.5	74.0	8400	93	29
	P			74	53	64	62			
17-21	$\bar{X}$	58.3	23.6	32.5	51.3	65.6	80.8	11,000	100	12
	P			88	75	79	80			
										105

*This figure is based on total sample for whom WRAT FK and income data were available, not traced group only. It is only approximate since data were grouped into intervals.

Table 9: Cohort III Blocked NPS-FT, Traced Math Scores

WRAT Math FK		PPV FK	PSI FK	MAT SK	MAT S1	MAT S2	MAT S3	Income*	% White	n
1-4	$\bar{X}$	40.5	10.0	24.4	37.0	50.4	68.4	8400	57	7
	P	44.7		59	36	18	42			
5-7	$\bar{X}$	48.1	15.4	19.9	34.8	49.7	60.1	8300	84	19
	P			31	28	16	18			
8-10	$\bar{X}$	50.2	18.1	21.3	37.0	49.8	62.6	7600	93	15
	P			40	36	16	25			
11-13	$\bar{X}$	54.8	19.5	22.7	38.5	53.5	65.3	8400	87	31
	P			50	41	31	31			
14-16	$\bar{X}$	55.1	21.3	28.8	44.8	60.2	73.1	9800	97	29
	P			76	56	62	60			
17-21	$\bar{X}$	59.0	24.0	34.0	55.5	68.1	82.6	9300	100	17
	P			90	86	83	83			
										118

*This figure is based on total sample for whom WRAT FK and income data were available, not traced group only. It is only approximate since data were grouped into intervals.

Table 10: Cohort III Blocked HS-FT, Traced Math Scores

WRAT Math FK		PPV FK	PSI FK	MAT SK	MAT S1	MAT S2	MAT S3	Income*	% White	n
1-4	$\bar{X}$	41.3	14.7	21.6	28.4	39.6	52.0	5000	29	7
	P			43	9	5	7			
5-7	$\bar{X}$	45.9	15.6	22.6	32.1	46.3	60.4	5700	56	16
	P			49	16	10	19			
8-10	$\bar{X}$	42.4	16.5	22.1	31.0	46.2	56.7	5300	40	20
	P			46	14	10	12			
11-13	$\bar{X}$	49.5	18.8	25.1	37.8	51.1	62.0	7100	61	23
	P			62	40	20	23			
14-16	$\bar{X}$	51.4	20.9	27.4	43.8	55.5	68.3	7100	65	31
	P			72	54	44	41			
17-21	$\bar{X}$	53.7	23.0	31.9	47.0	53.9	66.4	7400	57	7
	P			86	62	42	36			
										104

*This figure is based on total sample for whom WRAT FK and income data were available, not traced group only. It is only approximate since data were grouped into intervals.

Table 11: Cohort IV Blocked HS-NFT, Traced Reading Scores*

PSI Reading F HS		PPV F HS	PSI S HS	PPV S HS	MAT SK	MAT S1	MAT S2	Income*	% White	n
0-6	$\bar{X}$	23.7	15.9	36.4	23.4	35.2	45.7	31.0	36	10
	P				42	33	19			
7-9	$\bar{X}$	19.4	15.8	31.1	23.4	32.3	42.0	28.7	11	9
	P				42	24	11			
10-12	$\bar{X}$	26.4	15.7	40.7	24.5	36.3	46.1	35.4	39	13
	P				48	37	20			
13-15	$\bar{X}$	34.6	19.2	43.6	21.2	35.8	48.6	25.8	40	6
	P				36	35	31			
16-18	$\bar{X}$	34.2	18.8	42.4	25.0	39.0	50.3	33.2	50	6
	P				50	46	40			
19-32	$\bar{X}$	40.1	22.1	49.7	27.4	43.3	56.6	43.9	50	12
	P				72	67	68			
										56

*This figure is based on total sample for whom PSI F HS and income data were available, not traced group only. It is only approximate since data were grouped into intervals.

Table 12: Cohort IV Blocked HS-FT, Traced Reading Scores

PSI Reading F HS		PPV F HS	PSI S HS	PPV S HS	MAT SK	MAT S1	MAT S2	Income*	% White	n
0-6	$\bar{X}$	19.2	10.3	30.5	23.7	31.7	44.5	36.9	17	30
	P				44	23	15			
7-9	$\bar{X}$	22.7	14.7	36.0	23.9	33.0	44.5	34.6	29	49
	P				45	26	15			
10-12	$\bar{X}$	30.4	17.7	41.7	25.7	36.5	47.1	41.4	47	66
	P				56	38	23			
13-15	$\bar{X}$	32.7	18.6	44.1	27.0	38.2	49.1	39.6	41	51
	P				68	43	34			
16-18	$\bar{X}$	36.1	21.9	46.3	28.3	44.5	54.6	37.9	61	36
	P				80	72	58			
19-32	$\bar{X}$	47.0	26.5	53.6	29.7	47.0	59.5	46.1	88	41
	P				89	82	75			
										243

*This figure is based on total sample for whom PSI F HS and income data were available, not traced group only. It is only approximate since data were grouped into intervals.

Table 14: Cohort IV Blocked HS-FT, Traced Math Scores

PSI Math F HS		PPV F HS	PSI S HS	PPV S HS	MAT SK	MAT S1	MAT S2	Income*	% White	n
0-6	$\bar{X}$	18.8	9.9	30.2	20.6	32.3	46.3	33.6	19	26
	P				36	17	10			
7-9	$\bar{X}$	22.1	14.4	35.6	20.0	32.7	46.8	32.8	28	51
	P			32	19	11				
10-12	$\bar{X}$	30.0	17.6	41.6	22.7	37.8	49.8	40.3	44	63
	P				50	39	16			
13-15	$\bar{X}$	32.9	18.5	44.3	25.6	39.2	53.2	39.8	43	51
	P				64	43	29			
16-18	$\bar{X}$	36.5	21.7	46.1	27.1	43.5	56.0	37.9	56	34
	P				70	53	44			
19-32	$\bar{X}$	46.9	26.5	53.4	32.7	51.1	61.8	46.1	88	40
	P				89	74	69			
										264

*This figure is based on total sample for whom PSI F HS and income data were available, not traced group only. It is only approximate since data were grouped into intervals.

In summary, the most striking finding of our analyses is the longitudinal pattern of scores for different groups. It appears that Head Start experience does boost children's performance on tests used at fall of kindergarten and that Follow Through has a clear impact on maintaining Head Start graduates' high scores above those of Head Start gradutes who do not attend Follow Through kindergartens. Also, it seems clear that Follow Through boosts spring kindergarten scores of children who did not attend preschool.

V. Conclusion

In this section we present conclusions and recommendations suggested by our analyses. First we review the limitations of the Head Start-Follow Through data base, and how these restrict generalization. Second, we consider the three factors isolated in Section III as contributing to confounding in the Follow Through versus Non-Follow Through comparison. Using the data to specify more precisely how these factors have operated in our sample, we reach conclusions about the probable form of bias in the national evaluation. Finally, we consider some implications for future evaluations of programs like Head Start and Follow Through

Limitations of the Data

The data analyses presented above were exploratory, and may have stretched the data beyond the bounds that some would consider justifiable. In this study, we have sacrificed a certain degree of rigor in the hope of uncovering potentially important aspects of the data which might be missed

by more conservative attacks. Since our approach has several implications for the interpretation of national evaluation results, it is important to underline the limitations of the data analyzed here.

One limitation stems from the fact that HSPV is a special kind of Head Start program. Within any site, the Head Start and Follow Through programs were associated with the same sponsors and were based on the same theories of child development. Thus the relationship between Head Start and Follow Through for these children may differ in some way from that generally prevailing in the non-planned variation Head Start programs. On the other hand, reports by Smith (1973) and Weisberg (1974) on HSPV indicate that the effects of Head Start were quite uniform across models. On most measures, most Head Start programs seemed to show some effectiveness, with one or two exceptionally high or low.

Another limitation is the small number of sites with longitudinal data of any kind. Very few models and sets of demographic circumstances are represented. Inclusion of more sites may have produced somewhat different results. We suspect that the particular aspects we have uncovered in our analyses cut across sites and models, but we cannot be sure.

Further, since we believed it important to trace a consistent group at different time points, the sample was restricted further. In order to have adequate sample sizes we lumped together sites and sponsors, treating HS and FT as uniform programs.

We believe that the results presented are highly suggestive and plausible, despite these limitations. But we certainly cannot rule out the possibility that a more adequate data set would generate different results.

In fact, we believe that the methods employed in our analyses might fruitfully be used to explore the larger data base used in the national evaluation.

If feasible, it would be particularly interesting to apply the approach separately to sponsors, and even sites. Our methods are only suggestive, but they do not involve complex statistical artifacts of the sort involved in ANCOVA or regression analyses.[4] Thus, variation in such effects across sponsors and sites might tell us something important about the differential effectiveness of model/site combinations.

Implications for Confounding in the National Evaluation

In Section III we identified three intertwined factors which determine what is actually being estimated in the national evaluation. The more precisely we can specify how these factors have operated in Head Start and Follow Through, the more we can say about the adequacy of national evaluation results. We now discuss each of the factors in light of our analyses.

a) The proportions of children experiencing each of the four combinations of preschool and early elementary experience.

Throughout Abt's various reports (c.f. Stebbins et al., 1975), there is ample evidence that in most sites, there is a substantially higher

[4] It is only fair to point out that we began analyzing the HSPV-FT data set using analyses of covariance of the sort employed in the national evaluation, but found it virtually impossible to interpret differences in ANCOVA effects estimates of different grade levels and so switched to the more straightforward analyses reported here.

proportion of Head Start children in Follow Through than in the Non-Follow Through comparison groups. Follow Through aimed to recruit Head Start graduates, and seems to have been generally successful in doing so. Our special sample reflects this pattern. In Cohort III, there were virtually no Head Start-Non-Follow Through children. Cohort IV had a small Head Start-Non-Follow Through group.

According to the model presented in section III, the difference between the proportion p of Head Start children in Follow Through and q, of Head Start children in Non-Follow Through is a key factor. In estimating program effects at the end of third grade, the exact importance of this difference depends, however, on other factors. For now, let us note that in most comparisons in the national evaluation, there is reason to believe that

$$p > q$$

and hence that overall comparisons of FT and NFT children at the end of third grade will entail a bias tending artificially to underestimate FT effects.

b) Background differences among the streams of children flowing into each of the preschool-elementary groups.

In Cohort III, children in No Preschool-Follow Through and No Preschool-Non Follow Through seemed quite similar in terms of several indices reflective of academic potential. Children in Head Start-Follow Through seemed generally less "advantaged" than either of the No Preschool groups.

In our Cohort IV analysis, we compared HS-FT and HS-NFT children. As shown in section II, the selection process determining the composition of these groups may have varied considerably across sites. In some sites, children with higher growth potential tended to be in FT, and in others they tended to be in NFT. Thus the exact form of the S's in the model presented in section III may vary across sites.

In summary, in our samples, it seems that HS children tend to be more disadvantaged than NPS children. Conditional on being in NPS, selection into FT may not involve much additional bias. Conditional on being in HS, the nature of selection biases may be a local phenomenon.

All this suggests that, in terms of our model

$$S_{FT} = 0$$

$$S_{HS} < 0$$

$$S_{HS+FT} < 0$$

Moreover, it <u>may</u> be the case, at least approximately that

$$S_{HS+FT} = S_{HS}$$

Let us for convenience define

$$S_{HS+FT} = S_{HS} = S.$$

This model reflects the possibility suggested by our data that selection factors operate primarily through selection into Head Start.

Recall now that we defined the hypothetical WRAT kindergarten entry score which would have been obtained if Head Start did not exist by W. Since these scores would probably reflect the same forces as S, we can speculate that approximately

$$W_{HS+FT} = W_{HS} = W < 0$$

$$W_{FT} = 0$$

c) The influence of Head Start on entering kindergarten performance.

We have fairly strong evidence that Head Start has contributed to increased WRAT scores at fall kindergarten. All Head Start groups had higher percentiles at SK than one would expect for disadvantaged children. Moreover, for a given score on the WRAT, Head Start-Follow Through children in Cohort III performed lower on some other entering kindergarten measures than either the No Preschool-Follow Through or No Preschool-Non-Follow Through groups. This suggests that the WRAT may have been increased by Head Start participation. Thus some of the gains during HSPV reported by Smith (1973) and Weisberg (1974) have been maintained at least through the end of kindergarten.

In terms of the model developed above, it seems likely that the increment to the WRAT resulting from HS, which we have called α, generally satisfies

$\alpha > 0$

Taking our results here, and substituting in the last equation, we find that the quantity being estimated by an overall comparison between Follow Through and Non-Follow Through groups--some sort of Follow Through effect--can be represented as

$$\varepsilon = p\ (\tau_{HS+FT} + S - \beta W - \beta\alpha) + (1-p)\tau_{FT} - q\ (\tau_{HS} + S - \beta W - \beta\alpha)$$

To understand the nature of possible biases more precisely, let us consider the following special case:

$$\tau_{HS+FT} = \tau_{FT} = \tau$$

$$\tau_{HS} = 0$$

That is, the effect of Head Start alone on third grade outcomes is negligible, and the effect of Head Start and Follow Through together is about the same as that of Follow Through alone. Ideally, in this case, ε should be equal to τ. However, we find

$$\varepsilon = \tau + (p-q)\ (S - \beta W - \beta\alpha).$$

That is, we have a bias of the form

$$\text{Bias} = (p-q)\ (S - \beta W - \beta\alpha).$$

Let us consider β in more detail. Recall that

$$p - q > 0$$

$$S < 0$$

$$W < 0$$

$$\alpha > 0$$

Now β is computed from the data, and reflects a relationship between entering kindergarten WRAT and third grade MAT. Therefore β is bound to be positive.

Suppose first that β is very small, near zero. Then we have

$$\varepsilon \cong (p-q)\ S < 0$$

That is, the adjustment is negligible, and the bias is a function of p-q and S. As β increases, the adjustment becomes larger. Suppose β is large enough to exactly compensate for differences had the Head Start-free covariate W been available. In this case

$$S - \beta W = 0$$

and

$$\text{Bias} = (p-q)\ (-\beta\alpha) < 0.$$

In order for the adjustment to reduce the bias completely to zero, β would have to be large enough so that

$$S = \beta(W + \alpha)$$

Now β is computed from relationships between observed WRAT and MAT outcomes. Generally these are pooled within-group relationships from the FT and NFT groups being compared. The resulting β could conceivably be larger than that which would have been produced using the HS-free WRAT, but this seems unlikely. The boost in WRAT score for those in HS would probably decrease rather than increase β, by weakening the relationship between WRAT and the outcome.

The upshot of all these assumptions and analyses is that there may be a bias against FT built into the national evaluation. This bias would be a function of any difference in the proportions of HS children between FT and NFT groups being compared. Thus, it would be most serious when this difference is large.

Of course in this simplification we have ignored other covariates besides the WRAT which describe children prior to FT. A complex multivariate adjustment using more than one covariate may indeed be helpful if additional covariates can successfully untangle the components of the pretest attributable to HS from those attributable to background differences between groups. This would, however, require a very sophisticated interactive model, and the burden would be on the analysts to demonstrate its adequacy.

We would propose a simpler approach. Rather than stretching the linear adjustment approach beyond its limits, suppose we simply block, or stratify, on the basis of preschool experience. For children from HS, we have in terms of our model

$p = 1, q = 1 \Rightarrow \text{Bias} = 0$

For children with no PS, we have

$p = 0, q = 0 \Rightarrow \text{Bias} = 0$

Thus conducting analyses of covariance within blocks would eliminate the important sources of bias we have been considering.

Blocking also has a theoretical advantage. We do not need to worry about a possible dependence of FT effect on PS experience. Since τ_{FT} and $\tau_{HS+FT} - \tau_{HS}$ are being estimated separately, differences between them can be interpreted substantively. Without blocking, as we noted in section III, in overall comparisons between FT and NFT, the meaning of what is being estimated is somewhat ambiguous.

It is somewhat encouraging in this regard that, conditional on being in No Preschool, groups with and without Follow Through seemed very similar in terms of growth potential. Conditional on being in Head Start, results were less clear, but not discouraging. Thus it seems hopeful that at least for some sites or sponsors, blocking may remove the major problems in estimating a Follow Through effect.

A Rough Quantitative Estimate of the Bias in the National Evaluation

In the previous section, we presented a theoretical argument which suggested that estimates of effects in the national evaluation contain some negative bias. Particularly in sites with a much larger proportion of Head Start children in Follow Through than in Non-Follow Through, we would expect the relative effect of Follow Through to be underestimated. It is not at all clear, however, that the magnitude of this bias is large enough to cast serious doubts on the validity of the analyses. The pattern of effective and ineffective sites may be influenced very little by the forces we have considered. To explore this important issue further, we derive in this section a very rough estimate of the magnitude of bias.

In the previous section we derived an expression of the form

$$\text{Bias} = (p-q)\ (S-\beta W-\beta\alpha).$$

This expression is complex and difficult to evaluate precisely in a real situation. Note, however, that whatever the exact values of S, β, W, and α may be, the bias according to our model is proportional to (p-q), the difference between the proportions of Head Start children in Follow Through and Non-Follow Through. Now consider any site i, and let

$$B_i = (S_i-\beta W_i-\beta\alpha_i).$$

Then

$$\text{Bias}_i = (p_i-q_i)\ B_i.$$

If we define the effect estimated by ANCOVA for site i as Effect_i, then

$$\text{Effect}_i = (p_i-q_i)\,B_i + T_i$$

where T_i is the true effect. Now the relationship between B_i and p_i-q_i may be quite complex. Suppose for simplicity that the B_i's do not vary much relative to the (p_i-q_i)'s. Then

$$\text{Effect}_i = (p_i-q_i)B + T_i.$$

Under this simplistic model we can estimate B by a linear regression analysis with Effect as the dependent variable and p-q the independent variable. We can then obtain an estimate of the bias for any site by multiplying this estimate by p_i-q_i.

To perform the regression analysis, we have used the site-level effects estimates from the "local" analysis of Cohort III data described by Abt Associates (Stebbins et al., 1975). We have considered only the Total Reading and Total Math effects estimates for the 86 sites included in this analysis.[5]

The regression analysis yielded the following results:

$B_{\text{Reading}} = -.036 \qquad R^2 = .13$

$B_{\text{Math}} = -.075 \qquad R^2 = .23$

These coefficients have been translated into estimated biases for various values of p-q in Table 15.

---

[5] These data are drawn directly from the Abt Volume IVC report. Values for p and q were taken from tables entitled "Distribution of Background Variables" for each sponsor. Values of the estimated site effects were taken from tables entitled "Adjusted Differences between FT and NFT" for each sponsor. Eighty-six of the 87 cohort III-FK, III-EF and II-EF sites for which data are given in the Abt IVC volume are included here. The Oakland site is not included because no local comparison was available for that site.

Table 15: Estimated Biases Associated with Preschool Confounding

	Bias	
p-q	Reading	Math
.10	- .36	- .75
.20	- .72	-1.50
.30	-1.08	-2.25
.40	-1.44	-3.00
.50	-1.80	-3.75
1.00	-3.60	-7.50

Now as a criterion of "material significance" Abt used 1/4 of a standard deviation of the whole NFT population. For the two tests considered here

1/4 Total Reading Standard Deviation = 3.97

1/4 Total Math Standard Deviation = 4.41.

If an estimated effect exceeds this value, Abt has declared a positive effect (+). If it is less than minus this value, a negative effect (-) is indicated. Otherwise the effect is considered null (0).

Now for any site i, we have estimated a bias of size

$$\text{Bias}_i = (p_i - q_i)B$$

Thus by using the coefficients presented above, and the values of $p_i - q_i$, we obtain a numerical estimate of bias. Using this estimate we can "correct" the original Abt estimate. This procedure has been carried out for each of the 86 sites.

Both the original and the preschool corrected results are shown in Table 16. For each sponsor, the pattern of effects is displayed. Note first that on the basis of the totals for all 86 sites, the overall effect of Follow Through seems somewhat negative for math, and quite negative for reading. After correction, the math results seem less negative overall: they convey an impression of general equality between Follow Through and Non-Follow Through. Reading results remain largely unchanged.

Table 16: Patterns of Site Effects by Sponsor With and Without Preschool Correction

Sponsor		Total Math			Total Reading		
		-	0	+	-	0	+
Responsive Education	National analysis	5	1	3	4	3	2
	Preschool corrected	2	4	3	3	4	2
TEEM	National analysis	4	4	3	5	6	0
	Preschool corrected	3	6	2	5	6	0
Bank Street	National analysis	4	1	1	3	2	1
	Preschool corrected	3	2	1	2	3	1
Direct Instruction	National analysis	4	0	7	3	6	2
	Preschool corrected	3	1	7	3	6	2
Behavior Analysis	National analysis	1	3	3	1	6	0
	Preschool corrected	1	3	3	1	6	0
Cognitive Curriculum	National analysis	3	2	1	4	1	1
	Preschool corrected	1	4	1	4	1	1
Parent Education	National analysis	2	4	2	2	5	1
	Preschool corrected	1	5	2	1	6	1
EDC	National analysis	2	5	0	3	4	0
	Preschool corrected	2	4	1	3	4	0
IELP	National analysis	1	2	0	2	1	0
	Preschool corrected	1	2	0	2	1	0
ILM	National analysis	1	1	0	1	1	0
	Preschool corrected	1	1	0	1	1	0
SEDL	National analysis	0	4	1	3	1	1
	Preschool corrected	0	4	1	3	1	1
CLA	National analysis	0	2	0	1	1	0
	Preschool corrected	0	1	1	1	1	0
Home-School	National analysis	0	1	0	1	0	0
	Preschool corrected	0	1	0	1	0	0
Cal. Process	National analysis	2	1	0	2	1	0
	Preschool corrected	1	2	0	1	2	0
New School	National analysis	2	0	0	1	1	0
	Preschool corrected	2	0	0	1	1	0
Mathe-magenic	National analysis	0	1	2	2	1	0
	Preschool corrected	0	1	2	2	1	0
TOTAL	National analysis	31	32	23	38	40	8
	Preschool corrected	21	41	24	34	44	8

Looking at individual sponsors, we see that in terms of math, some have substantially changed. The Responsive Education and Cognitive Curriculum models, in particular, appear much improved. Three of the nine REC sites change from - to 0, substantially altering our conclusions about the model's performance. On the other hand, most models change little.

In conclusion, it seems fair to say that ignoring preschool experience may have changed results substantially for some models and some tests in the local analysis. Even if all the assumptions of the ANCOVA model were met, this factor has introduced some bias. However, it does not seem to us likely that this factor alone could play a dominant role in determining the patterns of effects reported in the National Evaluation. As indicated above, there are serious questions about the applicability of the ANCOVA model in analyzing quasi-experimental compensatory education data of this kind. If one accepts this model, however, then our analyses do not give evidence of major biases attributable primarily to the way preschool experience has been handled.

Implications for Future Studies

In this report we have tried to explicate the nature of problems in estimating Follow Through effects via the traditional approaches used in the national evaluation. In the course of this effort, however, we have become more firmly convinced that problems with the national evaluation lie at a deeper level. We feel that an evaluation approach is needed which looks not at average outcomes at some arbitrary point, such as third grade, but focuses rather on the ways in which individual growth has been influenced by an

intervention. Such an approach faces head on the conceptual problems inherent in attempts to affect children's development, rather than hoping for across the board gains so large and consistent that fine-grained analysis is unnecessary. We conclude this report with a brief discussion of some possible aspects of such an evaluation strategy.

First, it would be necessary to obtain data at several points in time for the same children. The longitudinal pattern of growth for a child contains much more information than does an end-of-program outcome. How the child reached a certain point may be predictive of future performance, which is ultimately more important than performance during the program.

An important feature of the analyses presented here is the tracing of an intact group across several time points. By observing the patterns over time, we uncovered a consistent effect of Head Start and Follow Through on performance through the end of kindergarten. This early boost may have important implications even though its effect on later measured skills is not evident in our data.

Coupled with the need for longitudinal data is the need for a longitudinal metric. Standardized tests are cross-sectional in that scores are based on the ranking of individuals relative to the norming sample. They do not provide an absolute measure of competence. Changes in score are therefore difficult to interpret. To assess effects of programs on individual children requires a metric on which individual change scores are meaningful. Traditional psychometric tests are constructed so as to render measurement of individual change virtually meaningless (c.f. Cronbach and Furby, 1970).

With multi-time point data collected on individuals using a longitudinal measure, it would be necessary to develop appropriate standards of comparison against which to compare individual patterns of performance over time. This would require a model of the "expected" performance of an individual with certain characteristics under one or more sets of standard conditions. The difference between the actual pattern of growth observed and that expected under standard conditions would provide easily interpretable information on the program's effect.

Of course the effect which can be observed during the program does not necessarily tell us much about future growth after the program is finished. Therefore it is necessarily to consider theories of development which may shed light on the meaning of short-term gains for future prospects.

Our data seem to indicate that both Head Start and Follow Through were boosting scores on tests taken in Head Start and throughout kindergarten. Test scores at later time points did not show evidence of similar gains. A pessimistic interpretation would be that early gains have faded out, so that Head Start and Follow Through have no long-term benefits.

In the context of specific developmental theories, these findings might support alternative interpretations. It might be possible to show that the levelling off observed for second and third grade still leaves these children better off than they would have been without any intervention. Perhaps they have indeed been raised to a slightly better developmental path. Better designed empirical studies would be helpful to pin down such a finding, but without some theoretical assumptions, the all-important extrapolation beyond the end of Follow Through is not possible.

Another important possibility is that the MAT tests are measuring skills which are not necessarily developmentally connected from kindergarten through third grade. This suggests the need to monitor children's progress across a wide range of developmental dimensions. Perhaps the early boost reflected in kindergarten WRAT and MAT scores is paying off in some other area, although it does not seem to show up in third grade MAT scores.

In summary, we have discussed here several components of an evaluation approach focusing not on estimates of average increment at a particular point, but rather on changes in patterns of individual growth. Such an approach is ambitious, but, we believe, would provide richer information on program performance. While not providing a simple index suitable for a "go - no go" decision, it would help decision makers by providing relevant information on the ways the program has functioned.

The components we have enumerated are:

1) Data on individuals across several time points
2) A longitudinal metric in terms of which individual changes are meaningful
3) Standards of comparison for individual patterns of growth
4) Theory of the developmental processes under study
5) Multi-dimensional data

At present, there is no simple blueprint available for implementing an evaluation designed along these lines. The development of methods to do this seems to us important if evaluations of innovative programs such as Head Start and Follow Through are to be more useful to decision-makers.

Perhaps the most important reason for adopting this kind of strategy is a very practical one: simplicity of communication. In this report, we have had to make many simplifications just to get a handle on how analyses of covariance may have performed under certain assumptions. The statistical technology employed in the national Follow Through evaluation is extremely complex and confusing, even to statistical experts.

Unfortunately, this complex technology is of such a special form that it is very difficult to integrate substantive information on programs into its workings. This has been a source of frustration to both Abt and the Follow Through sponsors, in their attempts to provide a fair picture of this program. The kind of evaluation we are suggesting would have the advantage that its assumptions would be open to inspection and argument by experts in child development and education who do not hold doctorates in mathematical statistics. We feel it is significant that nothing more sophisticated than means and percentiles have been used in our analyses. Since evaluation is far from a hard science, the assumptions and methodology employed in evaluation, as well as the conclusions reached, must be accessible to decision-makers.

Abt Associates. (Nov. 9, 1976). Draft. Education As Experimentation: A Planned Variation Model. Vol. IVC. Appendices. Cambridge, MA: Abt Associates, Inc.

Bryk, A.S. and Weisberg, H.I. (1977). Use of the Non-equivalent Control Group Design When subjects Are Growing, Psychological Bulletin (forthcoming).

Campbell, D.T. and Erlebacher, A. (1970). How Regression Artifacts in Quasi-Experimental Evaluations Can Mistakenly Make Compensatory Education Look Harmful. In J. Hellmuth (Ed.), Compensatory Education: A National Debate. Vol. III, 185-210, Disadvantaged Child. New York: Brunner-Mazel.

Campbell, D.T. and Frey, P.W. (1970). The Implications of Learning Theory for the Fade-Out of Gains from Compensatory Education. In J. Hellmuth (Ed.), Compensatory Education: A National Debate. Vol. III, Disadvantaged Child. New York: Brunner-Mazel.

Cicerelli, G., et al (1969). The Impact of Head Start: An Evaluation of the Effects of Head Start on Children's Cognitive and Affective Development. (Westinghouse Learning Corporation and Ohio University. Contract b89-4536 with the Office of Economic Opportunity) Washington, D.C.: Office of Economic Opportunity.

Cronbach, L.J. and Furby, L. (1970). How We should Measure "Change"--or Should We?, Psychological Bulletin, 74, 68-80.

Cronbach, L.J., Rogosa, D.R., Floden, R.E., and Price, G.G. (1976). Analysis of Covariance: Angel of Salvation or Temptress and Deluder? Occasional Paper, Stanford Evaluation Consortium.

David, J. (1973). Final Report on Sample Selection. Cambridge, MA: The Huron Institute.

Dittman, L.L. et al. (1971). Study of Selected Children in Head Start Planned Variation, 1969-70. Menlo Park, CA: Stanford Research Institute.

Durost, W.N., Bixler, H.H., Wrightstone, J.W., Prescott, G.A., Balow, I.H. (1971). Metropolitan Achievement Tests Teacher's Handbooks. New York: Harcourt, Brace Jovanovich.

Education Division, Office of the General Council, DHEW (1976). January 27, 1976 Memorandum to Mr. Fred Bresnick, Program Coordinator, Division of Follow Through, entitled "Follow Through Sponsors and the Buckley Amendment - Response to Your Memorandum dated August 26, 1975."

Elmore, R. (1976). Follow Through: Decision-Making in a Large Social Experiment, dissertation, Harvard Graduate School of Education.

Evans, J.W. and Schiller, J. (1970). How Preoccupation with Possible Regression Artifacts Can Lead to Faulty Strategies for the Evaluation of Social Action Programs: A Reply to Campbell and Erlebacher. In J. Hellmuth (Ed.), Compensatory Education: A National Debate. Vol. III, 216-225, The Disadvantaged Child. New York: Brunner-Mazel.

Magidson, J. (1977). Towards A Causal Model Approach for Adjusting Preexisting Differences in the Nonequivalent Control Group Situation: A General Alternative to ANCOVA. Cambridge, MA: Abt Associates, Inc.

Mendelsohn, R. (1970). Is Head Start a Failure or Success. In J. Hellmuth (Ed.), Compensatory Education: A National Debate. Vol. III, Disadvantaged Child. New York: Brunner-Mazel.

Smith, M.S. and Bissell, J.S. (1970). Report Analysis: the Impact of Head Start. Harvard Educational Review, Vol. 40, No. 1, pp. 51-104.

Smith, M.S. (1972). Head Start/Follow Through Planned Variations Longitudinal Study, memorandum to Lois-Ellin Datta and Garry McDaniels, The Huron Institute.

Smith, M.S. (1973). Some Short Term Effects of Project Head Start: A Preliminary Report on the Second Year of Planned Variation - 1970-71. Cambridge, MA: The Huron Institute.

Stebbins, L., et al (1975). Education as Experimentation: A Planned Variation Model. Vol. III. Cambridge, MA: Abt Associates.

Walker, D., Bane, M. and Bryk, A. (1973). The Quality of the Head Start Planned Variation Data. Cambridge, MA: Huron Institute.

Weisberg, H.I. (1974). Short Term Cognitive Effects of Head Start Programs: A Report on the Third Year of Planned Variation - 1971-1972. Cambridge, MA: The Huron Institite. ERIC number ED 093-047.

Wolff, M. and Stein, A. (1966). Six Months Later. A Comparison of Children Who Had Head Start, Summer 1965, with Their Classmates in Kindergarten. New York: Yeshiva University.

Table A-1. Cohort III, NPS-NFT Group

Site	Age	% Black	% White	Inc.	Mom Ed.	64-Item PSI FHS	64-Item PSI SHS	29-Item PSI FK	WRAT M-FK	WRAT R-FK	WRAT M-SK	WRAT R-SK	RW MAT M-SK	RW MAT R-SK	RW MAT M-S1	RW MAT R-S1	RW MAT M-S2	RW MAT R-S2	RW MAT M-S3	RW MAT R-S3	SS MAT M-SK	SS MAT R-SK	SS MAT M-S1	SS MAT R-S1	SS MAT M-S2	SS MAT R-S2	SS MAT M-S3	SS MAT R-S3
204	54.9	0.0	92.5					19.8	12.2	15.6	17.4	24.5	16.4	15.3	36.3	51.3	61.5	60.2	59.8	55.0	27.0	27.2	40.2	42.9	54.6	55.5	68.6	62.6
	4.3							3.6	3.4	6.1	2.1	3.7	6.5	4.6	8.5	15.0	17.8	15.4	19.2	15.3	6.3	3.5	7.3	8.3	10.3	8.2	10.3	8.9
	67	67	67					63	64	64	50	50	49	50	51	51	63	63	33	54	49	50	51	51	63	63	53	54
213	55.1	8.3	79.2					20.8	12.3	17.4	16.3	22.6	15.4	14.5	41.7	50.3	66.0	59.3	53.3	53.1	25.8	26.6	45.1	43.7	57.6	55.5	64.5	62.9
	3.9							4.1	3.9	7.1	2.8	5.2	6.4	4.5	7.9	20.2	16.7	19.1	20.3	19.7	6.9	3.6	8.2	12.8	8.7	11.3	12.2	14.1
	48	48	48					38	36	36	37	37	43	43	18	18	29	25	36	36	43	43	18	18	29	25	36	36
316	54.0	.9	97.1					20.6	12.9	17.3	16.1	23.6	16.5	15.4	40.1	54.9	65.1	50.9	66.1	61.7	27.2	27.1	44.3	45.6	57.8	56.5	71.7	67.2
	3.5							4.2	3.9	6.8	3.2	6.6	7.2	6.6	10.7	18.1	17.9	18.7	20.9	19.2	8.0	6.2	10.8	12.2	10.0	10.5	11.9	13.7
	110	110	110					104	104	104	64	64	64	64	92	92	50	50	68	69	64	64	92	92	50	50	68	69
804	53.6	6.3	93.8					17.8	9.1	8.3	15.6	19.9	13.4	13.4	41.0	53.2	52.6	52.3	54.0	46.0	24.1	25.5	46.5	44.6	50.2	52.3	64.9	56.7
	1.9							4.9	5.5	6.5	3.1	6.0	4.5	4.5	9.0	13.7	16.3	17.6	19.3	18.1	4.4	3.8	9.1	10.0	9.4	9.7	11.2	12.3
	16	16	16					13	15	15	15	15	15	15	13	13	15	15	15	15	15	15	13	13	15	15	15	15
902	51.8	17.0	83.0					16.2	8.7	11.8	12.5	17.9	8.8	9.6	32.7	42.0	56.2	49.3	55.7	47.4	18.2	21.7	37.6	37.7	52.9	50.8	66.4	58.1
	4.2							4.2	4.3	5.1	3.6	7.5	4.0	5.5	10.7	14.3	17.7	18.7	19.6	18.5	5.6	5.1	7.6	7.5	10.1	9.4	11.1	10.6
	47	47	47					43	43	43	46	46	40	44	37	38	41	41	37	38	40	44	37	38	41	41	37	38
906																												
910																												

Table A-2. Cohort III, NPS-FT Group

Site	Age	% Black	% White	Inc.	Mom Ed.	64-Item PSI FHS	64-Item PSI SHS	29-Item PSI FK	WRAT M-FK	WRAT R-FK	WRAT M-SK	WRAT R-SK	RW MAT M-SK	RW MAT R-SK	RW MAT M-S1	RW MAT R-S1	RW MAT M-S2	RW MAT R-S2	RW MAT M-S3	RW MAT R-S3	SS MAT M-SK	SS MAT R-SK	SS MAT M-S1	SS MAT R-S1	SS MAT M-S2	SS MAT R-S2	SS MAT M-S3	SS MAT R-S3
204	33.6	0.0	97.1					19.7	10.3	15.6	16.1	23.4	13.0	13.0	36.9	47.3	63.1	57.6	55.5	53.7	22.3	25.3	40.7	40.4	55.9	54.0	66.3	67.4
	3.6							2.9	3.4	6.1	2.8	5.8	8.1	5.1	8.6	18.5	14.6	18.4	16.0	17.5	8.7	5.1	7.1	10.6	7.5	8.5	6.5	15.6
	34	34	34					33	31	31	32	32	27	28	30	31	31	31	24	24	27	28	30	31	31	31	24	24
213	35.4	26.3	63.2					21.3	15.2	21.5			14.4	17.6	30.2	42.0	59.8	45.4	43.4	43.3	24.9	29.1	36.2	38.5	53.6	49.4	58.4	56.3
	3.7							3.5	2.1	7.0			6.5	6.8	13.8	15.5	15.7	18.1	18.8	20.6	6.1	6.6	10.7	9.2	9.1	9.2	12.4	12.9
	19	19	19					10	10	10			14	14	17	16	17	16	12	12	14	14	17	16	13	16	12	12
316	33.9	6.5	92.0					19.2	11.8	16.9	16.3	25.2	13.6	17.0	34.5	48.8	58.2	49.2	57.6	50.3	24.1	28.5	39.7	41.8	53.8	51.0	67.4	66.7
	3.9							4.3	4.5	6.4	3.7	5.8	6.1	5.5	12.0	17.6	19.0	21.2	21.0	19.9	6.5	4.9	10.8	10.5	10.7	12.2	12.0	13.5
	137	137	137					127	127	127	126	126	126	126	111	110	99	103	90	90	126	126	111	110	99	103	90	90
504	34.5	13.5	85.2					16.7	10.1	12.6	18.3	24.5	20.4	17.1	41.2	50.1	74.1	56.8	75.9	63.8	31.6	28.9	46.7	43.0	62.3	55.2	77.8	66.5
	2.8							5.5	5.3	6.6	4.0	8.3	8.7	7.0	15.0	21.6	17.8	21.5	26.2	17.0	10.1	6.3	15.8	14.6	10.2	11.2	15.3	16.2
	19	19	19					18	18	18	18	18	16	16	19	19	16	16	14	13	16	16	19	19	16	16	14	13
902																												
906	33.7	0.0	58.3										14.6	12.8	27.8	36.4	61.3	50.3			24.8	25.1	34.1	34.4	56.0	53.0		
	5.6												6.9	3.3	14.9	14.5	22.6	20.0			5.8	2.6	11.1	7.7	13.9	13.6		
	12	12	12										12	12	11	11	11	11			12	12	11	11	11	11		
910	33.3	42.9	28.6					12.6	8.1	15.4	14.5	24.4	12.0	15.8	28.5	36.3	39.7	41.3	56.9	45.0	22.6	27.5	34.2	35.2	42.5	44.8	66.6	57.3
	4.0							5.6	5.2	9.2	-.7	4.0	4.4	3.4	5.2	17.3	12.6	24.7	26.8	16.-	5.0	2.2	3.0	9.7	9.3	13.0	15.5	10.2
	14	14	14					10	10	10	10	10	12	12	12	13	11	13	10	10	12	12	12	12	13	13	10	10

Table A-3. Cohort III, HS-NFT Group

Site	Age	% Black	% White	Inc.	Mom Ed.	64-Item PSI FPS	64-Item PSI SnS	29-Item PSI FK	WRAT M-FK	WRAT R-FK	WRAT M-SK	WRAT R-SK	RW MAT M-SK	RW MAT R-SK	RW MAT M-S1	RW MAT R-S1	RW MAT M-S2	RW MAT R-S2	RW MAT M-S3	RW MAT R-S3	SS MAT M-SK	SS MAT R-SK	SS MAT M-S1	SS MAT R-S1	SS MAT M-S2	SS MAT R-S2	SS MAT M-S3	SS MAT R-S3
204																												
213																												
316	54.0	28.6	66.7	40.2	11.2	53.1	51.1	10.4	10.4	15.0	15.2	18.9	11.9	12.8	25.5	33.1	48.7	51.0			22.3	24.8	34.9	36.4	48.2	49.8		
	3.9			22.8	1.8	14.1	12.5	6.4	4.4	7.0	2.6	6.6	4.4	4.6	9.8	14.6	17.2	15.0			5.2	4.0	7.5	7.2	9.2	6.9		
	21	21	21	21	10	13	13	20	18	18	16	16	15	15	10	10	10	11			15	15	10	5	10	11		
804																												
902																												
906	53.5	0.0	46.2	40.9	9.6	36.7	41.3				16.3	24.5	11.7	16.6							22.5	27.9						
	3.4			24.3	2.2	7.0	8.7				2.0	5.1	3.9	4.3							3.8	2.8						
	13	13	13	13	13	10	12				12	12	12	12							12	12						
910																												

Table A-4. Cohort III, HS-FT Group

Site	Age	% Black	% White	Inc.	Mom. Ed.	64-Item PSI FBS	64-Item PSI SBS	29-Item PSI FK	WRAT M-FK	WRAT R-FK	WRAT M-SK	WRAT R-SK	RM MAT M-SK	RM MAT R-SK	RM MAT M-S1	RM MAT R-S1	RM MAT M-S2	RM MAT R-S2	RM MAT M-S3	RM MAT R-S3	SS MAT M-SK	SS MAT R-SK	SS MAT M-S1	SS MAT R-S1	SS MAT M-S2	SS MAT R-S2	SS MAT M-S3	SS MAT R-S3
204	55.1 3.1 72	6.9 72	80.6 72	61.4 26.3 61	10.8 2.0 24	37.4 12.3 27	48.1 8.9 27	19.6 4.1 69	12.3 3.9 69	14.5 5.7 69	16.5 2.9 65	22.3 4.6 65	12.4 5.3 60	13.5 4.1 64	34.8 10.4 50	42.0 15.0 51	58.1 16.1 36	52.0 18.1 38	55.7 17.9 26	49.9 19.8 28	28.7 6.4 60	25.4 4.3 64	39.2 10.0 50	37.8 8.1 51	53.2 8.6 36	52.0 8.8 38	65.9 9.9 26	59.5 12.2 28
213	55.3 3.7 51	27.5 51	64.7 51	39.2 20.5 51	10.9 2.1 47	37.5 9.2 19	44.3 9.9 21	19.5 4.5 49	12.3 4.2 49	14.4 5.6 49	14.9 3.9 23	20.1 7.4 23	14.0 5.9 42	12.4 4.9 43	25.5 11.2 29	31.3 10.4 29	43.5 16.3 31	33.9 13.5 32	35.8 18.0 18	34.0 12.5 23	24.4 6.1 42	24.2 5.5 43	32.2 8.4 29	31.9 6.3 29	44.6 10.9 31	42.8 7.3 32	51.7 16.1 16	50.9 7.2 23
316	53.4 5.0 45	13.3 45	71.1 45	44.2 25.1 45	11.0 1.6 43	31.5 9.9 20	45.1 7.7 20	16.7 4.3 45	10.0 4.0 45	13.0 6.1 45	13.8 3.7 40	20.5 5.9 40	11.3 5.5 39	12.9 4.9 40	28.8 10.7 32	38.3 15.7 31	48.7 18.0 25	42.4 17.8 34	49.6 19.4 23	39.6 13.4 24	21.5 6.0 39	25.1 4.4 40	34.6 8.3 32	35.8 9.2 31	47.4 10.6 25	50.5 7.0 34	62.5 11.2 24	54.1 8.0 24
804	53.9 3.6 53	41.5 53	58.5 53	35.5 13.3 53	9.6 2.0 53	31.3 12.0 53	44.8 9.8 52	16.9 5.4 53	11.2 4.0 52	14.6 5.2 52	17.7 4.2 51	23.8 5.2 51	19.2 6.2 49	15.8 4.1 49	36.8 12.5 44	41.9 14.4 44	60.0 16.4 34	30.7 11.2 26	66.3 14.6 28	50.6 13.6 28	29.4 6.4 49	27.2 3.0 49	41.5 11.8 44	37.7 7.6 44	54.5 9.0 34	40.7 6.7 26	68.7 7.9 28	60.1 7.2 28
902	51.5 4.4 38	89.5 38	10.5 38	34.3 18.6 33	10.0 2.1 38	21.4 8.5 35	38.2 9.2 36	15.4 4.5 38	7.4 3.5 37	8.9 2.0 37	12.5 3.2 37	14.9 4.6 37	8.7 2.4 34	9.4 4.0 36	14.8 7.3 24	25.3 9.5 27	38.8 12.7 26	55.3 17.5 13	36.1 12.3 25	29.5 13.3 25	19.0 3.4 34	21.8 .40 36	23.9 6.6 24	27.6 6.5 27	41.5 8.4 26	53.8 10.0 13	54.2 8.6 25	46.4 10.6 25
906	54.9 3.6 32	0.0 32	34.4 32	37.1 13.7 32	9.4 2.0 32	42.9 9.9 26	44.5 10.1 26	18.7 5.4 30	20.8 4.7 29	12.2 4.9 29	15.5 3.9 28	21.3 5.5 28	13.8 5.8 29	14.7 4.8 29	28.8 10.9 20	37.9 14.5 20	57.1 19.3 13				24.2 5.6 29	26.3 3.7 29	34.3 8.0 20	36.4 10.7 20	53.1 10.3 13			
910																												

Table A-[illegible] Cohort IV, NPS-NET Group

Site	Age	% Black	% White	Inc.	Mom Ed.	30-Item PSI FHS	30-Item PSI SHS	RW MAT M-SK	RW MAT R-SK	RW MAT M-S1	RW MAT R-S1	RW MAT M-S2	RW MAT R-S2	SS MAT M-SK	SS MAT R-SK	SS MAT M-S1	SS MAT R-S1	SS MAT M-S2	SS MAT R-S2
204	54.6	1.5	96.9					14.2	16.3	38.2	53.2	61.5	61.9	24.8	27.7	41.9	43.7	55.2	56.9
	4.2							5.7	[illegible]	9.5	16.0	18.3	16.8	5.4	4.7	8.5	9.0	9.9	9.3
	196	196	196					109	113	80	87	61	62	109	113	80	87	61	62
209	54.2	.8	86.2					13.4	16.6	42.9	55.5	57.2	57.6	23.8	38.1	46.8	46.3	52.5	54.7
	4.2							5.6	5.9	10.7	18.3	18.2	19.5	5.8	5.8	11.2	12.3	10.4	11.0
	123	123	123					163	163	99	88	83	90	163	163	99	88	83	90
309	54.9	37.6	60.8					13.8	16.1	35.3	42.5	56.9	48.5	24.3	27.6	40.5	38.5	53.3	50.7
	3.8							5.6	5.3	12.1	18.7	18.6	21.5	5.5	4.6	10.8	11.4	10.7	11.5
	125	125	125					107	110	73	74	72	76	107	110	73	74	72	76
316	54.0	1.1	97.4					14.5	17.4	36.0	52.2	63.3	60.0	25.3	28.6	40.9	44.0	56.8	56.5
	3.5							6.5	5.9	12.6	18.7	20.2	19.0	7.2	5.7	11.4	11.8	12.1	11.8
	352	352	352					326	331	260	256	221	222	326	331	260	256	221	222
511	51.6	84.8	11.6					11.1	12.9	32.3	38.5	52.8	42.5	23.6	25.0	37.9	35.8	50.5	46.5
	3.9							4.6	4.7	12.5	16.3	16.2	17.9	5.0	4.2	11.0	9.9	9.0	8.3
	302	302	302					226	238	141	149	135	141	226	238	141	149	135	141
804	54.8	21.7	78.3					10.9	12.5	39.0	49.8	61.2	62.0	21.1	24.6	43.3	42.4	55.5	56.7
	4.1							4.4	4.5	13.7	20.9	13.3	17.2	5.7	4.0	12.0	11.9	5.9	10.1
	59	60	60					43	44	38	36	2[illegible]	23	43	44	38	36	24	23
808																			
902	52.6	40.0	60.0					8.0	10.7	27.1	34.0	51.6	40.0	17.6	23.0	33.5	33.2	50.4	46.2
	3.8							3.0	4.4	10.0	10.4	12.[illegible]	15.7	4.9	4.2	7.3	4.8	6.1	7.3
	63	65	65					49	59	48	48	3[illegible]	37	49	59	48	48	37	37
905	54.6	0.0	60.6					13.1	15.5	39.8	46.1	60.[illegible]	54.2	23.8	26.9	43.6	39.8	54.4	53.2
	3.8							4.9	6.2	10.3	17.9	17.6	18.0	4.9	5.1	9.2	10.4	9.9	10.3
	66	66	66					55	55	36	36	29	29	55	55	36	36	29	29
1203	51.3	0.0	100					14.5	16.8	38.6	52.6	65.3	60.2	25.3	38.1	42.6	43.4	57.9	55.6
	4.0							5.6	5.2	10.2	15.7	17.9	18.0	6.2	4.3	9.7	8.6	10.5	9.6
	137	137	137					130	129	96	98	12[illegible]	99	130	129	96	98	100	99
1204	55.9	0.0	98.9					16.5	16.9	45.7	50.5	71.[illegible]	63.6	27.2	28.4	51.7	43.8	60.5	58.9
	4.5							6.3	5.7	11.9	16.1	17.1	17.4	6.8	5.2	14.5	11.5	10.4	11.4
	179	179	179					150	153	70	71	122	123	150	153	70	71	122	123

Table A-6. Cohort IV, NPS-FT Group

Site	Age	% Black	% White	Inc.	Mom Ed.	32-Item PSI FHS	32-Item PSI SHS	RW MAT M-SK	RW MAT R-SK	RW MAT M-S1	RW MAT R-S1	RW MAT M-S2	RW MAT R-S2	SS MAT M-SK	SS MAT R-SK	SS MAT M-S1	SS MAT R-S1	SS MAT M-S2	SS MAT R-S2
20-	55.9	1.0	95.0					13.4	15.1	36.1	48.0			23.7	26.7	40.1	41.5		
	4.6							6.0	6.1	9.2	15.5			6.5	5.9	7.6	9.3		
	100	100	100					81	89	51	51			81	89	51	51		
209	54.3	5.6	58.3					11.7	14.1	36.1	39.7	54.7	44.1	21.7	25.0	40.3	36.9	51.5	47.8
	3.9							7.7	6.7	11.2	13.6	17.5	17.1	8.9	7.0	9.0	7.9	10.3	7.0
	35	36	36					26	27	18	18	11	12	26	27	18	18	11	12
213	54.6	18.6	67.8					10.3	12.7	25.4	34.6	43.2	35.5	20.0	24.8	32.1	33.6	44.2	42.5
	4.8							5.7	5.8	12.6	12.2	20.3	18.0	6.8	5.0	9.5	6.7	13.9	10.0
	38	59	59					43	49	27	28	12	11	43	49	27	28	12	11
309	52.5	20.7	59.2					7.4	9.4	20.9	43.7	46.3	44.1	17.0	22.1	37.8	39.1	46.5	49.1
	4.9							2.1	3.1	13.8	20.0	20.6	20.1	4.7	3.5	15.8	12.6	13.8	11.3
	169	169	169					5	8	39	39	27	27	5	8	39	39	27	27
316	53.8	6.0	90.0					14.0	17.7	34.4	49.8	63.1	54.2	24.7	29.0	39.3	42.8	55.8	53.6
	4.0							6.1	5.9	11.9	19.1	18.6	20.1	6.4	5.3	9.8	12.5	10.3	11.5
	299	299	299					252	255	187	186	108	108	252	255	187	186	108	108
511	50	100	0					10.7	14.0	26.7	33.8	38.1	30.7	20.8	25.9	33.4	33.0	42.1	41.2
	3.3							5.0	5.3	11.3	14.1	8.9	11.3	6.1	5.4	8.6	8.3	7.5	7.1
	24	24	24					20	23	21	21	13	13	20	23	21	21	13	13
804	54.9	11.4	88.6					19.2	18.4	39.6	47.9	66.4	56.0	29.3	29.0	44.8	41.8	57.6	52.9
	3.9							7.9	4.8	13.0	18.4	16.4	21.3	8.4	3.6	13.0	11.4	8.5	11.3
	44	44	44					38	39	31	30	11	11	38	39	31	30	11	11
808	52.4	18.4	81.6					15.2	16.3	41.8	54.8	72.0	60.8	25.6	27.6	46.3	45.4	62.2	55.9
	3.9							6.6	4.7	12.5	17.2	19.8	18.5	6.6	3.4	12.1	10.9	11.4	10.2
	37	38	38					30	30	25	25	21	21	30	30	25	25	21	21
902	50.7	66.7	29.2					6.6	8.3	18.4	25.0			15.4	20.6	27.1	27.6		
	5.5							2.5	5.0	9.7	9.5			5.2	4.7	8.1	5.4		
	48	48	48					33	34	21	21			33	34	21	21		
906	56.1	0	43.1					14.9	15.0	36.9	40.8			25.4	26.9	40.9	37.5		
	4.7							6.2	5.9	10.2	15.8			5.6	5.3	8.5	9.0		
	51	51	51					38	39	26	38			38	39	26	28		
1203	51.7	0	100.0					10.8	10.7	28.3	42.6			20.6	22.8	36.0	37.3		
	5.2							6.2	5.0	17.4	21.7			7.6	4.6	15.9	11.6		
	14	14	14					12	12	10	10			12	12	10	10		
1204	55.8	0	100.0					21.6	20.9	47.2	65.6	76.2	73.4	32.1	31.8	52.3	52.8	64.0	67.5
	4.7							6.3	5.1	9.9	12.8	16.9	13.5	7.0	5.2	12.1	10.5	11.0	12.9
	69	69	69					61	62	54	55	52	52	61	62	54	55	52	52

Table A-7. Cohort IV, HS-NF Group

Site	Age	% Black	% White	Inc.	Mom Ed.	32-Item PSI FHS	32-Item PSI SHS	RW MAT M-SK	RW MAT R-SK	RW MAT M-S1	RW MAT R-S1	RW MAT M-S2	RW MAT R-S2	SS MAT M-SK	SS MAT R-SK	SS MAT M-S1	SS MAT R-S1	SS MAT M-S2	SS MAT R-S2
204																			
209																			
213																			
309																			
316	54.4	0	100	36.4	11.1	14.8	20.4	12.0	14.7					22.9	26.6				
	4.2			18.8	1.6	6.6	6.3	3.7	5.5					4.0	4.5				
	18	18	18	18	18	18	16	15	15					15	15				
511																			
804	53.9	73.5	26.5	29.5	9.4	11.6	16.2	10.0	11.8	34.9	42.6	59.5	51.8	20.0	23.5	40.0	37.9	54.6	52.7
	3.7			18.3	2.9	4.0	5.1	5.6	5.7	13.9	18.0	17.2	17.9	7.3	6.5	13.3	10.3	9.1	10.1
	34	34	34	34	34	34	33	25	25	31	31	22	23	25	25	31	31	22	23
808																			
902	51.5	60.0	40	34	9.9	11.9	14.6	8.8	11.0	25.1	33.8	52.1	41.6	18.2	23.3	31.8	34.2	50.1	46.9
	3.1			13.	2.1	6.1	4.3	4.8	4.9	9.0	12.9	14.3	17.3	6.1	4.4	7.1	11.1	8.1	8.0
	20	20	20	20	19	18	18	13	19	17	17	13	13	13	19	17	17	13	13
906	54.9	0	20	47.3	8.7	13.3	16.5	10.6	10.6	29.4	31.0	51.1	36.8	21.1	22.8	35.1	31.3	47.6	44.5
	3.3			18.2	3.5	5.5	6.9	4.0	4.3	9.7	9.8	16.7	13.2	4.4	3.8	6.8	7.3	9.8	6.4
	20	20	20	19	19	19	18	18	18	16	15	16	15	18	18	16	15	16	15
1203																			
1204																			

Table A-8. Cohort IV, HS-FT Group

Site	Age	% Black	% White	Inc.	Mom Ed.	32-Item PST FHS	32-Item PSI SHS	RW MAT M-SK	RW MAT R-Sk	RW MAT M-S1	RW MAT R-S1	RW MAT M-S2	RW MAT R-S2	SS MAT M-SK	SS MAT R-SK	SS MAT M-S1	SS MAT R-S1	SS MAT M-S2	SS MAT R-S2
104	54.9	7.9	89.5	37.0	10.9	15.4	21.8	11.8	15.3	35.7	46.0	59.4	52.0	22.5	26.8	39.7	39.9	54.3	52.5
	3.5			14.0	1.3	5.7	5.5	4.8	5.2	9.9	14.2	15.8	18.9	4.7	4.0	8.1	7.5	8.6	9.8
	38	38	38	37	24	36	36	33	35	21	21	23	24	33	35	21	21	23	24
209	55.9	3.6	53.6	32.5	10.9	14.8	21.3	13.3	15.1	36.3	45.7	58.5	49.7	23.8	26.6	40.4	39.5	52.5	50.1
	4.1			14.5	1.5	5.3	4.5	4.4	5.9	11.6	16.4	18.8	20.0	5.4	4.0	9.3	8.6	11.4	8.7
	28	28	28	28	22	25	24	18	20	15	15	11	12	18	20	15	15	11	12
213	54.8	23.5	74.5	34.8	11.1	13.1	19.7	12.7	13.3	27.7	33.1	48.8	35.1	23.1	25.3	34.3	32.5	47.7	45.0
	4.0			11.1	1.6	5.8	5.4	6.0	4.5	12.4	14.2	18.5	14.7	6.5	4.1	9.6	8.4	10.7	8.2
	51	51	51	49	51	48	47	42	41	32	34	24	24	42	41	32	34	24	24
309	53.8	51.6	21.0	59.2	10.4	11.3	15.3	10.8	13.8	31.4	35.3	45.9	39.0	21.0	25.8	36.4	33.6	46.5	45.7
	3.3			23.9	3.2	5.5	6.0	5.1	4.2	9.6	14.6	17.3	17.2	6.0	3.5	7.2	8.5	10.7	8.8
	62	62	62	36	55	48	53	42	41	49	51	41	40	42	41	49	51	41	40
316	53.5	17.1	74.3	35.8	10.2	13.9	19.0	11.5	15.4	30.4	43.9	52.0	46.7	21.9	27.2	35.8	38.6	49.9	49.2
	3.7			13.8	1.7	5.5	5.5	4.9	5.2	11.5	15.0	16.8	18.9	5.4	4.5	9.1	7.4	9.6	9.8
	35	35	35	34	34	33	32	30	31	23	21	20	20	30	31	23	21	20	20
511	51.6	98.3	1.7	31.7	10.6	8.6	13.0	10.9	11.8	29.0	32.3	46.7	36.9	21.6	24.1	34.6	32.2	47.0	44.0
	3.4			17.1	1.6	4.3	4.5	3.6	3.8	9.9	10.2	14.8	14.8	3.7	3.4	7.2	6.0	9.5	7.7
	60	60	60	59	58	59	53	52	55	44	45	42	44	52	55	44	45	42	44
804	55.1	40	60	37.7	9.6	13.1	21.2	20.3	18.5	42.5	51.6	68.9	56.9	31.0	29.4	47.0	43.9	60.6	54.3
	3.5			20.3	2.8	6.4	6.6	7.7	5.1	11.2	16.2	20.0	18.5	9.4	4.3	11.6	10.9	12.9	4.8
	55	55	55	55	55	55	50	50	50	43	43	35	35	50	50	43	43	35	35
808	52.4	80	20	44.5	10.1	13.1	15.0	13.4	16.3	35.2	46.5	59.4	49.4	23.4	27.6	39.5	40.3	54.7	50.6
	3.1			20.6	1.8	5.3	6.8	7.2	5.1	11.0	14.9	18.3	16.9	8.8	4.2	9.0	7.7	10.1	8.8
	45	45	45	45	43	45	43	43	43	37	36	33	34	43	43	37	36	33	34
902	53.3	96.7	3.3	29.4	10.1	10.2	14.3	7.2	9.7	21.1	26.3	40.8	29.3	16.5	21.9	28.6	28.5	43.5	40.6
	3.3			14.8	1.7	4.8	5.3	2.4	4.2	9.3	6.2	11.2	8.1	4.0	4.1	7.1	4.7	8.1	7.1
	30	30	30	30	30	27	30	29	29	27	27	24	24	29	29	27	27	24	24
906	55.3	0	33.3	37.1	10.4	13.5	18.8	11.3	13.5	27.8	30.6	55.5	44.0	22.2	25.6	33.6	31.7	51.3	48.1
	3.3			14.6	2.7	5.4	4.8	3.6	5.7	10.2	9.1	23.0	18.2	3.8	5.6	8.1	6.0	14.7	8.9
	24	24	24	24	21	24	22	19	21	11	12	15	14	19	21	11	12	15	14
1203	51.2	0	100	41.3	11.0	11.4	18.3	11.6	11.9	26.1	37.9	50.9	47.7	21.9	24.0	32.6	35.6	49.5	49.8
	3.5			17.8	8.4	4.9	5.5	5.0	4.5	11.2	14.1	16.6	19.2	5.8	4.8	8.6	8.0	9.6	9.3
	65	65	65	57	65	64	59	57	56	49	50	38	39	57	56	49	50	38	39
1204	57.1	0	100	52.2	11.6	19.6	24.7	20.0	18.7	44.6	59.6	68.5	67.1	31.2	29.4	50.0	49.1	59.0	61.7
	4.9			19.8	1.0	5.9	3.8	7.4	3.7	12.0	13.5	21.3	16.0	9.3	2.7	13.7	11.2	13.3	13.8
	27	27	27	26	22	27	27	23	23	23	23	24	24	23	23	23	23	24	24

Appendix B: Merger of HSPV and FT Test Data

The merging of the separate HSPV and FT data sets was accomplished as follows. Huron already possessed two data tapes containing fall and spring test data for HSPV cohorts II and III (identified as CRIS and MASTER respectively). From Stanford Research Institute we obtained three data tapes containing all available data for Follow Through and Non-Follow Through tested children in each of the HSPV-FT sites (seven sites for CIII and twelve for CIV). The CIV data was transmitted on one tape (ID No. HU6916). The CIII data were contained on two separate tapes; one containing data for CIII children whose parents had been interviewed (ID No. HU 6423); and one containing data for children with no parent interview data (ID No. HU 4496). Since none of the CIV children had had parent interviews this distinction did not pertain to them.

Also from the Stanford Research Institute we obtained a copy of a matched list of identification numbers for children who were in both HSPV-FT. Since data gathering for HSPV and FT had been carried out separately, individual children had different ID numbers in the two programs. These were matched by hand by SRI. (This matching process is described in a letter dated 3 July 1974 from Richard Marciano to Dr. Marvin Cline; Subject: Transmittal of Head Start Planned Variation/Follow Through Children Match Data.) In merging the HSPV and FT data sets we used the standards set out in this letter as our criteria for satisfactory match. Specifically HSPV and FT data records were considered satisfactorily matched if within project locations the records had a good name match (either names completely the same or a near miss), plus there was a match on at least two of the following three criteria; sex, birthdate, and ethnicity).

Using these standards for matching the HSPV and FT sets were merged and recorded onto two magnetic tapes (identified as SSHFC3 and SSHFC4 for cohorts III and IV respectively). These tapes are described in separate codebooks, and it was from these worktapes that virtually all of the analyses reported above were conducted.